Roger Penrose Collected Works

VOLUME 2

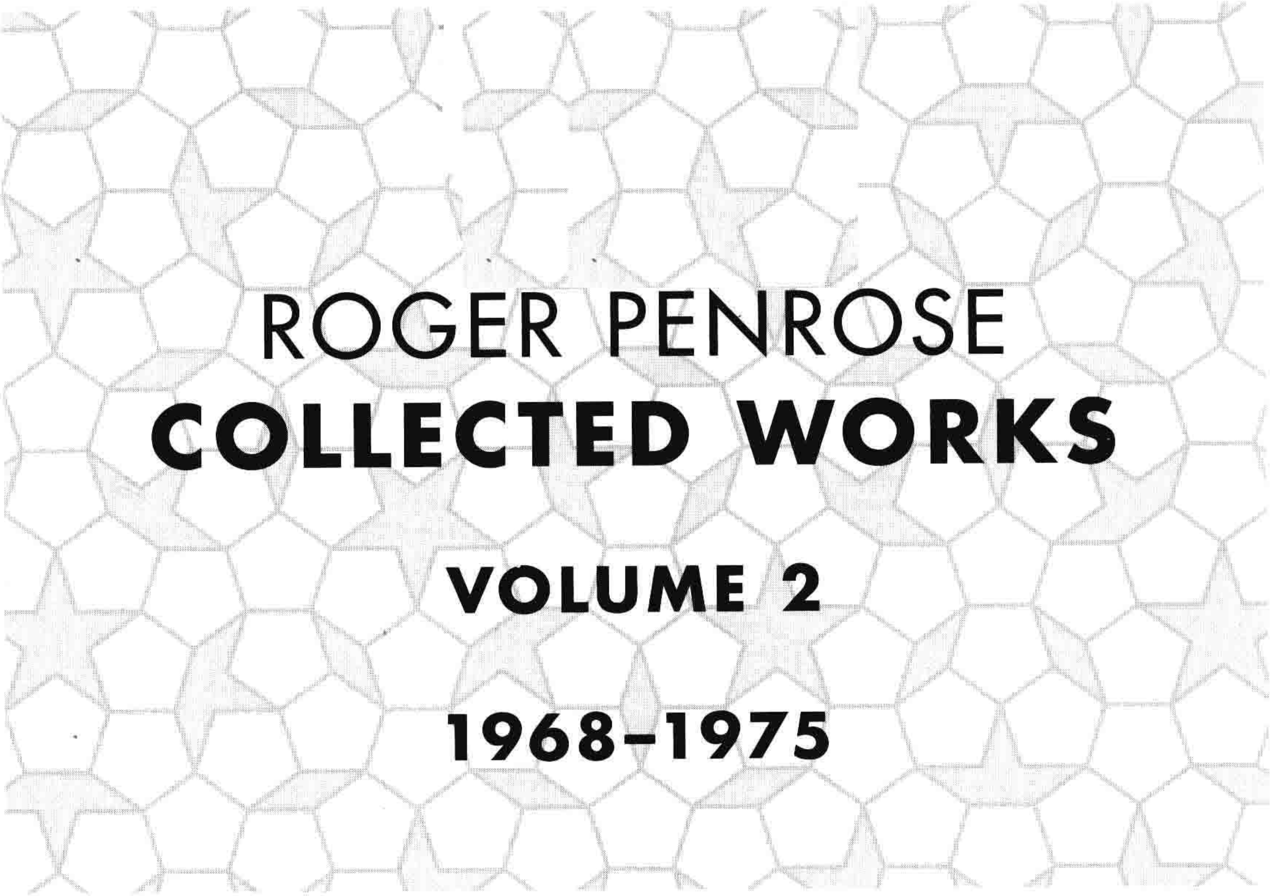
1968–1975

OXFORD SCIENCE PUBLICATIONS

OXFORD
UNIVERSITY PRESS

www.oup.com

ISBN 978 0 19 921937 7



ROGER PENROSE

COLLECTED WORKS

VOLUME 2

1968–1975

常州大学图书馆
藏书章

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

© Roger Penrose 2011

The moral rights of the authors have been asserted

Crown copyright material is reproduced under Class Licence
Number C01P0000148 with the permission of OPSI
and the Queen's Printer for Scotland

Database right Oxford University Press (maker)

First published 2011

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organization. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2010926298

Typeset by Newgen Imaging Systems Pvt Ltd., Chennai, India
Printed in Great Britain
on acid-free paper by
CPI Antony Rowe, Chippenham, Wiltshire

ISBN 978-0-19-921937-7 (Vol. 2)
978-0-19-921944-5 (Set)

1 3 5 7 9 10 8 6 4 2

PREFACE

The topics addressed, in these *Collected Works*, vary over quite an extensive range, the boundaries between their areas being sometimes quite blurred. It would have been hard to classify all these articles and arrange them according to individual topics. Moreover, my own mode of working has been, on the whole, to keep most of my different scientific or mathematical interests running concurrently. Accordingly, it has seemed appropriate not to attempt to group the papers together into separate tidy categories. To do so would have provided a somewhat misleading picture of my overall mode of thinking. Instead, the ordering of articles that has been adopted is essentially chronological, at least with regard to the date of publication. The resulting somewhat higgledy-piggledy appearance that we find in the sometimes abrupt juxtaposition of different topics is, perhaps, fairly illustrative of my actual mode of working.

There are two main kinds of drive that underlie these works. On the one hand, some of the articles have arisen mainly from purely mathematical considerations, which can be ultimately aesthetic or just matters of simple curiosity. On the other hand, motivations for many others of the articles came primarily from a strong desire to discover more about the underlying principles governing the workings of the physical world. Yet, one of the deep mysteries about what has been discovered about these principles is the great breadth of intrinsically appealing mathematics that is found to underlie the actual operation of the physics of the universe, so these two drives are found to relate closely to one another. The reader may perhaps note the distinct geometrical flavour of much of my own mathematics, evidencing the strong appeal that geometry exerts on me, an appeal that is not infrequently intertwined with desire for algebraic elegance. Indeed, I have found it remarkable how much of the workings of our universe appear to depend upon sophisticated and powerful geometry and algebra.

My own very geometrical perspectives, as opposed to purely computational ones, are perhaps made more obvious in the later volumes of this series than in the early ones, but this appearance is deceptive. Very early in my career as a Cambridge graduate student, I had developed a graphical notation which enabled me to carry out many kinds of calculation in an entirely diagrammatic way. For a long time I was rather secretive of this procedure, and in my earlier writings would take care to remove all trace of these visual origins. I do not doubt that my predilections for such geometric or algebraically elegant forms of understanding have influenced the specific choices of direction in which my endeavours towards furthering our understandings of the physics of the world have taken.

The focus of these directions has become more clarified as time has moved on, and has narrowed down to such a degree that there has emerged a somewhat disturbing repetitiveness in many of the later articles. A good number of these provide rather non-technical accounts. An example of such repetitiveness occurs with my descriptions of gravitational collapse to a black hole. There is, however, a gradually developing outlook on this issue, leading to an emphasis on the fundamental contrast in character between the space-time singularities that inevitably appear within black holes and the space-time singularity that constitutes the big bang, this contrast being central to the existence of a second law of thermodynamics in our universe. The same kind of repetitiveness applies to my frequently repeated claim of a need for a fundamental change in the basic laws of quantum theory, where only towards the end do my suggestions become particularly specific. Likewise, many mildly differing accounts are provided outlining the formalism and the basic ideas that underlie twistor theory. The later ones lead up to various viewpoints on the 'googly' problem, which has constituted the main obstruction to the finding of a full understanding of general relativity in twistor terms. Despite such repetition, it is my hope that there is a sufficient variety of emphasis and of historical development in such articles for it to have been worth while to include them all here.

Several colleagues have collaborated with me in writing the works presented here, and I can mention but only a few. Foremost, I have collaborated long and particularly fruitfully with Ezra T. Newman, especially with regard to the asymptotics of general relativity, and also with Wolfgang Rindler in the development of the formalism of 2-spinor calculus. Other collaborators who have played important roles in helping to develop the mathematics of discrete physics and twistor theory have been Erwin Kronheimer, Michael Eastwood, Ronny Wells, and Denny Hill. Among my collaborators, providing very significant inputs, who were former students in my research group, are Andrew Hodges, Paul Tod, Lane Hughston, George Sparling, Richard Ward, Nick Woodhouse, and Lionel Mason, but there were many others. Although not playing roles as actual collaborators, fundamental influences in my research have come from Dennis Sciama and Michael Atiyah.

Roger Penrose
April 2010

BIOGRAPHY

Early years and influences

I was born in Colchester, Essex, UK on 8 August 1931. My father, Lionel Sharples Penrose, FRS, was a highly talented and intelligent man, with many interests. He, himself, had come from a Quaker family, his own father (James Doyle Penrose) having been a professional artist and his mother (Josephine Peckover) was from a well-to-do Quaker banking family. Lionel's primary professional interest was in the causes of hereditary mental disease. He was Galton Professor of Human Genetics at University College London from 1945 and, following retirement, Director of the Kennedy–Galton Institue, St Albans. He had a very considerable artistic talent and a keen interest in music, especially Mozart, but also Bach and other classical composers, dabbling a little in composition himself, particularly when there was a puzzle aspect in what was involved (where he once wrote a piece that could be played equally well upside down as right-way up). He liked making puzzles of various kinds in wood, and spent much of his later life designing wooden models illustrating self-reproduction. He was recognized as a fine composer of chess problems. He had a definite fondness for mathematics and was an innovative statistician. He loved explaining matters of science and mathematics. On one occasion, when I told him that my mathematics teacher at school had informed us that he would be explaining calculus the next day, Lionel immediately took me aside so as to be sure to be the first to teach me about the beauties of calculus. This made a great impression on me. Yet he was an extremely difficult person to communicate with on personal matters.

My mother, born Margaret Leathes, was a charming and caring woman with a very subtle mind. She had been Head Girl at Bedales school, and became medically qualified, but found she was not able to practice after her marriage. After Lionel died, in 1972, she married Max Newman FRS, a fascinating man, who had been Professor of Mathematics in Manchester and, as we later learned, a key figure at Bletchley Park during the war. Margaret's father, John Beresford Leathes, FRS, had been a distinguished professor of physiology, and her mother, a concert pianist, born Sara Mara Natanson from a Jewish family in Latvia, had disowned her family (then in St Petersberg), becoming completely secretive about her origins, and even about her maiden name. Margaret enjoyed geometry at school, but did not communicate much on this topic with me. My older brother Oliver was very precocious, and I learnt a lot of physics and mathematics from him as I was growing up. He later became a distinguished mathematical physicist (statistical mechanics) and FRS. My younger brother Jonathan excelled at games, particularly chess, and later became British Chess Champion a record 10 times, and then a top correspondence chess player. My much younger sister Shirley Victoria, now Shirley Hodgson, became a highly regarded Professor of Cancer Genetics.

This atmosphere of science, mathematics, art, and music, and of puzzles and games, had been an important part of my upbringing. During the war years, the family lived in London, Ontario, Canada, where Lionel became Director of Psychiatric Research at the Ontario Hospital. We returned to England in 1945, where I attended University College School, London. As an undergraduate, I studied Mathematics at University College London, from 1951 to 1954; then I went to St John's College Cambridge to do research in algebraic geometry, initially under William Hodge and subsequently John Todd. In my same year, under Hodge, was Micheal Atiyah, who was an inspirational, though somewhat intimidating colleague. Frank Adams and Christopher Zeemen (see Chapter 13) were contemporaries who also influenced me, and I became good friends with Hans Liebeck and Douglas Munn (see Chapter 3), who were close research-student colleagues at St John's.

Yet, the colleague who had the greatest influence on the development of my research during this period, and for many years later, was the cosmologist Dennis Sciama. Even before studying at Cambridge, I had a chance meeting with him during a visit to my brother Oliver in Cambridge in which I raised a question that had worried me in relation to the inspiring series of BBC radio talks given by Fred Hoyle in 1951 (The nature of the universe. A series of broadcast lectures). On the strength of this encounter (at which I used a space-time diagram with tilted light cones to make my point), Dennis had evidently formed the opinion that it would be worth while to interest me more in physics, even trying, later, to persuade me to switch my topic of research to cosmology. I was then too committed to my many pure-mathematical activities, but I learnt a great deal of physics from Dennis, and of the excitements involved in striving to uncover secrets of the universe. Another influential colleague was Felix Pirani, from whom I learnt much about developments in general relativity.

Pivotal were the superb lecture courses of Bondi, on general relativity, and Dirac, on quantum mechanics. On the pure-mathematical side, I recall particularly an elegant information-packed algebra course by Philip Hall. Lectures by Steen on mathematical logic turned out to influence me strongly towards my later viewpoint concerning the relation between computation and conscious understanding (e.g. Chapters 169, 184, 186).

Degrees

1952	BSc Mathematics (Spec) (1st Cl), University College, London
1957	PhD (Cantab), St John's College, Cambridge

Employment

1956	NRDC London (research on numerical analysis with computers); later consultant
1956–1957	Assistant Lecturer, Bedford College, London
1957–1960	Research Fellow, St John's College, Cambridge
1959–1961	NATO Research Fellow, Princeton, Syracuse, Cornell Universities
1961–1963	Research Associate, King's College, London
1963–1964	Visiting Associate Professor, University of Texas, Austin, Texas
1964–1966	Reader (Applied Mathematics), Birkbeck College, London
1966–1967	Visiting appointments: Yeshiva, Princeton, Cornell, Chicago Universities; lecturer Battelle Inst., Seattle
1967–1973	Professor (Applied Mathematics), Birkbeck College, London
1969–1970	Visiting professor, Yeshiva New York, Princeton, and Cornell Universities
1973–1998	Rouse Ball Professor (Mathematics), Oxford; 1998– Emeritus
1978	Miller Fellow, Univ. of California, Berkeley
1982–1987	Lovett Professor (Mathematics), Rice Univ., Houston, Texas, USA (Part-time)
1987–1993	Distinguished Professor (Physics, Mathematics), Syracuse Univ. USA (Part-time)
1993–	Francis and Helen Pentz Distinguished Professor (Physics, Mathematics; part-time)
1998–2001	Gresham Professor (Geometry), Gresham College, London

Honours Prizes include 1971 Heinemann Prize; 1972 Fellow of Royal Society; 1978 Plenary Lecture, International Congress of Mathematicians, Helsinki; 1985 Royal Medal (Royal Society);1988 Wolf Prize (physics, with Hawking); 1990 Albert Einstein Medal; 1994 knighted (services to science); 1994 Polish Academy of Sciences; 1998 US National Academy of Sciences; 2000 Order of Merit; 2004 de Morgan Medal (London Mathematical Society); 2001 Royal Irish Academy (Science); 2008 Copley Medal (Royal Society).

Marriages

1958	Joan Isabel Wedge; divorced 1980
1988	Vanessa Dee Thomas

INTRODUCTION

In this volume, many of the ideas introduced in Volume 1 are developed further in various ways. Twistor theory, which in Volume 1 was presented as, basically, giving us just a geometrical framework for the representation of flat 4-dimensional space-time (Chapter 29), now becomes more concerned with genuine issues of physics. It gives us a way of generating (arbitrary analytic) massless fields in terms of free holomorphic twistor functions (Chapters 37 and 39); basic scattering processes involving massless fields are considered (Chapters 37 and 52); and we have the beginnings of twistor diagram theory (the twistor version of Feynman diagrams) and curved-space twistor theory (Chapters 52 and 69). Certain properties of black holes are uncovered (Chapters 43 and 54); cosmic censorship proposed (Chapters 40 and 60); and it is shown that, in principle, rotational energy of a black hole can be extracted (Chapters 40 and 47). The singularity theorems are also extended (Chapter 42). General procedures are given for identifying singular points or points at infinity for a space-time manifold (Chapter 50); the role of a spacelike conformal infinity is pointed out in relation to the value of the cosmological constant and to cosmological horizons (Chapter 33). The algebraic notion of 'abstract indices' is presented (Chapter 36), whereby the convenient 'physicist's' tensor notation can be reinterpreted in a pure-mathematically respectable and completely coordinate-free way, and this is related to a parallel diagrammatic notation (Chapter 46).

A thoroughly different kind of development is the identification of a set of six geometrical shapes that can tile the entire Euclidean plane only non-periodically, where the tiling necessarily contains arbitrarily large regions of exact 5-fold symmetry (Chapter 64).

ACKNOWLEDGEMENTS

Oxford University Press has attempted to trace the copyright holder of all the papers included in these volumes and apologizes to copyright holders if permission to publish in this form has not been obtained.

The author and publishers thank the following individuals for their permissions:

Acta Physics Polonica B **280**
Adam Hilger Ltd **147**
American Chemical Society **206**
American Institute of Physics
  American Journal of Physics **19, 24**
  Journal of Maths and Physics **15, 26, 27, 29, 39, 41, 58, 84**
  Errata **15**
American Mathematical Society **30, 180**
  Bulletin of the American Mathematical Society **144**
  Contemporary Mathematics **179**
American Physical Society
  Physical Review Letters **16, 20, 23, 302**
  Reviews of Modern Physics **21**
Annals of Human Genetics **100**
Asian Journal of Mathematics **281**
Astronomische Gesellschaft **295**

Benjamin **36**
Bibliopolis **175**
Blackwell **80, 166, 186, 215**
  Annals of the New York Academy of Sciences **60, 190, 294**
British Broadcasting Company **59**
British Journal of Psychology **9**

Cambridge Entomological Club **243**
Cambridge University Press **6, 7, 46, 104, 138, 157, 172, 173, 176, 195, 218, 220, 227, 230, 240, 290, 303, 304**
  Behavioural and Brain Sciences **193, 194, 224**
  Proceedings of the Cambridge Philosophical Society **3, 4, 5, 32, 42, 50, 126**
Chapman and Hall/CRC Press **107, 124, 125, 127, 128, 130, 136, 139, 142, 143, 146, 149, 150, 154, 155, 156, 160, 161, 164, 170, 171, 178, 189, 191, 198, 201, 203, 209, 229, 234, 293**
Connected Globe **257**
Consigle Nazionale Della Richerche **145**
Cornell University Press **33**

Éditions du Centre National de la Recherche Scientifique **14**
Elsevier **12, 45, 52, 68, 85, 86, 92, 93, 123, 151, 167, 185, 187, 226, 279**
Escher, M.C. images: The M.C. Escher Company B.V. **218, 249, 306**
Eureka **1, 2**

Folio Society **284**
Freeman **56**

Granta **297**
Gresham College **277**

Halsted Press **148**
Hanser **66, 67**

Imperial College Press **264**
Imprint Academic **235, 245**
Indian Academy of Sciences **244**
Indiana University Press **25**
Institute of Electrical and Electronics Engineers **196**
Institute of Physics **217**
  Classical and Quantum Gravity **253, 254, 274, 282**
  European Journal of Physics **258**
  Journal of Physics A: Mathematical and General **95**
International Astronomical Union
  International Astronomical Union Symposium Proceedings Series No. 13, Cosmology **72**

International Atomic Energy Agency
  Contemporary Physics **38**
International Centre for Theoretical Physics **53, 68**
International Mathematical Union **286**
International University Bremen **305**
Inter-University Centre for Astronomy and Physics **265**
The Italian Astronomical Society
  Memorie della Società Astronomica Italiana **200**

Longman **78, 79, 108, 116, 117, 134, 135, 141, 177, 188**

The Massachusetts Institute of Technology Press **228, 246**

National Institute of Sciences of Ukraine **48**
Nature **8, 44, 49**
Nature (Physical Science) **47**
New Scientist **10, 101**
New York Academy of Sciences **294**
Oxford University Press **55, 69, 121, 122, 165, 178, 184, 197, 202, 208, 216, 237, 270, 272, 276, 291, 299**

Pearson **109**
Penguin **273**
Penrose's Unpublished Papers **31**
Pergamon Press **17**
Pitman **73, 75, 76, 77, 82, 87, 89, 90, 94, 102, 103, 110**
Pontificial University **296**
Princeton University Press **28, 268**
  Annals of Math **13, 131**
Proceedings of the International Congress of Mathematicians **115**

Rivista del Nuovo Cimento **40**
Routledge **169**
Royal Astronomical Society **183**
The Royal Society
  Proceedings of the Royal Society Series A **22, 35, 97, 105, 106, 132**
  Philosophical Transactions of the Royal Society of London. Series A **275, 283**

- The Sciences **99**
- *Scientific American* **51**
- Simon and Schuster **242**
- Springer **61, 63, 65, 83, 91, 98, 111, 133, 168, 204, 207, 231, 255, 263, 278, 292, 298, 300**
  - *Artificial Intelligence Review* **205**
  - *Communications in Mathematical Physics* **43, 54, 129**
  - *Cosmology and Gravitation* **120**
  - *Foundations of Physics* **140**
  - *General Relativity and Gravitation* **70, 71, 113, 114, 239, 251**
  - *Higher-Order and Symbolic Computation* **287**
  - *International Journal of Theoretical Physics* **37, 56**
  - *Journal of Statistical Physics* **233**
  - *Lecture Notes in Physics* **112, 152**
  - *The Mathematical Intelligencer* **285**
- Taylor and Francis **18**
  - *International Studies in the Philosophy of Science* **256**
  - *Surveys in High Energy Physics* **119**
- Trinity College, Dublin **219**

- World Scientific **162, 163, 225, 236, 241, 252, 289, 301**

- Universal Academy Press **266**
- University of Chicago Press **96, 269**
- University of Chittagong **158**
- University College London **267**

- The University of Oxford (Mathematics Department)
  - *Twistor Newsletter* **73, 74, 75, 76, 77, 78, 79, 81, 82, 88, 89, 90, 94, 102, 103, 107, 108, 116, 117, 118, 124, 125, 127, 128, 130, 134, 135, 136, 137, 139, 141, 142, 143, 146, 149, 150, 153, 154, 155, 156, 157, 160, 161, 164, 170, 171, 174, 181, 182, 188, 189, 191, 192, 198, 199, 201, 203, 209, 210, 211, 212, 213, 214, 221, 222, 223, 229, 232, 238, 248, 290, 250, 259, 260, 261, 262, 271, 288**

- University of Utah Press **247**

Children

1963	Christopher Shaun
1964	Toby Nicholas
1966	Eric Alexander
2000	Maxwell Sebastian

Honorary degrees

1992	University of New Brunswick, Canada
1993	University of Surrey
1994	University of Bath
1995	University of London
1996	University of Glasgow
1996	University of Essex
1997	University of St Andrews
1998	University of Santiniketan, India
1998	Open University
2002	University of Southampton
2004	University of Waterloo, Canada
2005	University of Leuven, Belgium
2005	Warsaw University
2005	University of Athens
2006	University of York

Books

1972	Techniques of Differential Topology in Relativity (NSF and SIAM)
1984 and 1986	Spinors and Space-Time (Volumes 1 and 2; Cambridge University Press; with Rindler)
1989	The Emperor's New Mind (Oxford University Press 1989)
1994	Shadows of the Mind (Oxford University Press)
1996	The Nature of Space and Time (Princeton University Press; with Hawking)
1997	The Large, the Small and the Human Mind (Cambridge University Press)
2004	The Road to Reality (Jonathan Cape)

CONTENTS
VOLUME 2

Introduction *xxiii*

1960s

33 Cosmological boundary conditions for zero rest-mass fields 1
34 On gravitational collapse and cosmology 15
35 New conservation laws for zero rest-mass fields in asymptotically flat space-time 29
36 Structure of space-time 61
37 Twistor quantization and curved space-time 177
38 On gravitational collapse 217
39 Solutions of the zero-rest-mass equations 231
40 Gravitational collapse: The role of general relativity 235
41 Conserved quantities in the Einstein–Maxwell theory 261

1970s

42 The singularities of gravitational collapse and cosmology 267
43 On quadratic first integrals of the geodesic equations for type {22} spacetimes 289
44 Scattering of two impulsive gravitational plane waves 301
45 Applications of negative dimensional tensors 307
46 Angular momentum: An approach to combinatorial space-time 333
47 Extraction of rotational energy from a black hole 367
48 Spinor classification of energy tensors 375
49 Black holes and gravitational theory 389
50 Ideal points in space-time 395
51 Black holes 419
52 Twistor theory: An approach to the quantisation of fields and space-time 429
53 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants' 505
54 On a quadratic first integral for the charged particle orbits in the charged Kerr solution 513
55 The geometry of impulsive gravitational waves 521

56 On the nature of quantum geometry 537
57 Internal instability in a Reissner–Nordström black hole 583
58 A space-time calculus based on pairs of null directions 599
59 Black holes 609
60 Naked singularities 635
61 Relativistic symmetry groups 647
62 Gravitational collapse 707
63 Gravitational collapse 729
64 The role of aesthetics in pure and applied mathematical research 741
65 Singularities in cosmology 749
66 Twistors and particles: An outline 761
67 The non-linear graviton 779
68 Classical gravity: A status report 791
69 Twistor theory: Its aims and achievements 795

DETAILED CONTENTS
VOLUMES 1–6

VOLUME 1

1950s

1 Morley's trisector theorem 1
2 'Construction problem' and 'Word problem' 3
3 A note on inverse semigroups 5
4 A generalized inverse for matrices 11
5 On best approximate solutions of linear matrix equations 21
6 Tensor methods in algebraic geometry 25
7 Geometrical algebras: A new approach to invariant theory 251
8 A self-reproducing analogue 351
9 Impossible objects: A special type of visual illusion 355
10 Puzzles for Christmas 359
11 The apparent shape of a relativistically moving sphere 365

1960s

12 A spinor approach to general relativity 369
13 Imbedding of manifolds in Euclidean space 403
14 General relativity in spinor form 415
15 An approach to gravitational radiation by a method of spin coefficients 421
16 Asymptotic properties of fields and space-times 437
17 The light cone at infinity 441
18 Conformal treatment of infinity 447
19 Energy conservation as the basis of relativistic mechanics 469
20 Gravitational collapse and space-time singularities 475
21 A remarkable property of plane waves in general relativity 479
22 Zero rest-mass fields including gravitation: Asymptotic behaviour 487
23 10 exact gravitationally conserved quantities 533
24 Energy conservation as the basis of relativistic mechanics II 537
25 General-relativistic energy flux and elementary optics 541
26 Note on the Bondi-Metzner-Sachs group 559
27 Some new gravitationally conserved quantities 569
28 An analysis of the structure of space-time 579
29 Twistor algebra 731
30 Conserved quantities and conformal structure in general relativity 755
31 Theory of quantized directions 769
32 On the structure of causal spaces 801

VOLUME 2

1960s

33 Cosmological boundary conditions for zero rest-mass fields 1
34 On gravitational collapse and cosmology 15
35 New conservation laws for zero rest-mass fields in asymptotically flat space-time 29
36 Structure of space-time 61
37 Twistor quantization and curved space-time 177
38 On gravitational collapse 217

39 Solutions of the zero-rest-mass equations 231
40 Gravitational collapse: The role of general relativity 235
41 Conserved quantities in the Einstein–Maxwell theory 261

1970s

42 The singularities of gravitational collapse and cosmology 267
43 On quadratic first integrals of the geodesic equations for type {22} spacetimes 289
44 Scattering of two impulsive gravitational plane waves 301
45 Applications of negative dimensional tensors 307
46 Angular momentum: An approach to combinatorial space-time 333
47 Extraction of rotational energy from a black hole 367
48 Spinor classification of energy tensors 375
49 Black holes and gravitational theory 389
50 Ideal points in space-time 395
51 Black holes 419
52 Twistor theory: An approach to the quantisation of fields and space-time 429
53 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants' 505
54 On a quadratic first integral for the charged particle orbits in the charged Kerr solution 513
55 The geometry of impulsive gravitational waves 521
56 On the nature of quantum geometry 537
57 Internal instability in a Reissner–Nordström black hole 583
58 A space-time calculus based on pairs of null directions 599
59 Black holes 609
60 Naked singularities 635
61 Relativistic symmetry groups 647
62 Gravitational collapse 707
63 Gravitational collapse 729
64 The role of aesthetics in pure and applied mathematical research 741
65 Singularities in cosmology 749
66 Twistors and particles: An outline 761
67 The non-linear graviton 779
68 Classical gravity: A status report 791
69 Twistor theory: Its aims and achievements 795

VOLUME 3

1970s

70 Nonlinear gravitons and curved twistor theory 1
71 The nonlinear graviton 25
72 The space-time singularities of cosmology and in black holes 33
73 The twistor quadrille 45
74 The snail contour 53
75 The universal bracket factor 55
76 Some elementary twistor integrals with boundary 59
77 Twistor functions and sheaf cohomology 63
78 How to grow (or at least half grow) new twistors from old 77
79 Some remarks on the two twistor description of electromagnetic and linearized gravitational potentials 81
80 The cosmic censorship hypothesis 83
81 A finite covering for PT^+ 87
82 The back-handed photon (or, to a cricketer, the googly photon!) 89
83 Any space-time has a plane wave as a limit 99
84 The Kähler structure of asymptotic twistor space 105
85 The twistor programme 113
86 Space-time singularities 127
87 The twisted camel (or how to get through the eye of a singularity). Includes: more on the twisted camel 137
88 Remarks concerning the ambidextrous photon 145
89 Massless fields and sheaf cohomology 149
90 Note on the n-twistor internal symmetry group 159
91 Aspects of general relativity 163
92 Is nature complex? 203
93 Gravitational collapse: A review 213
94 Local H^1's and propagation 231

95 Extended Regge trajectories 237
96 Singularities of spacetime 243
97 The metric and curvature properties of $\mathcal{H}$-space 271
98 The geometry of the universe 297
99 Einstein's vision and the mathematics of the natural world 341
100 The topology of ridge systems 347
101 Twisting round space-time 359
102 A googly graviton? 365
103 Twistors as helicity raising operators 375
104 Singularities and time-asymmetry 381
105 General introduction 441
106 Global general relativity 445
107 Remarks on the Sparling and Eguchi-Hanson (googly?) gravitons 449
108 The anti-self-dual Coulomb field's non-Hausdorff twistor space 459
109 On the twistor descriptions of massless fields 461
110 Combinatorial quantum theory and quantized directions 499
111 Pentaplexity: A class of nonperiodic tilings of the plane 507

1980s

112 Recent advances in global general relativity: A brief survey 515
113 Twistors for flat and curved space-time 527
114 Null hypersurface initial data for classical fields of arbitrary spin and for general relativity 575
115 The complex geometry of the natural world 617
116 On the evaluation of twistor cohomology classes 625
117 Three channels for the box diagram 629
118 Puzzle: A sequence for twistorians 637
119 A brief introduction to twistors 639
120 A brief outline of twistor theory 663

VOLUME 4

1980s

121 Time-asymmetry and quantum gravity 1
122 Some remarks on twistors and curved-space quantization 31
123 On Schwarzschild causality—a problem for "Lorentz covariant" general relativity 47
124 Concerning a Fourier contour integral 61
125 A new angle on the googly graviton 65
126 A quadratic mapping with invariant cubic curve 73
127 Twistors for cosmological models 91
128 Physical left-right symmetry and googlies 97
129 Cohomology and massless fields 105
130 A theory of 2-surface ('superficial') twistors 153
131 Some unsolved problems in classical general relativity 161
132 Quasi-local mass and angular momentum in general relativity 201
133 Time-asymmetry, cosmological uniformity and space-time singularities 213
134 Real-space topology for advanced/retarded twistor functions 217
135 Twistor blisters and relative cohomology 221
136 On the geometry of googly maps 225
137 A puzzle for twistorians 229
138 Some remarks on gravity and quantum mechanics 231
139 General-relativistic kinematics? 241
140 Spinors and torsion in general relativity 247
141 Non-locality in physics: A new mathematical viewpoint 263
142 On closed-set coverings for PT$^+$—A correction 269
143 Remarks on curved-space, twistor theory and googlies 271
144 A note on Sparling's 3-form 277
145 Physical space-time and nonrealizable CR-structures 281
146 Quasi-local mass and angular momentum 305
147 Relative cohomology, googlies and deformations of $\mathbb{I}$ 309
148 Donaldson's moduli space: A 'model' for quantum gravity? 313
149 Mass in general relativity 319
150 'New improved' quasi-local mass, and the Schwarzschild solution 331
151 On Bryant's condition for holomorphic curves in CR space 339

152 Integrals for general-relativistic sources: A development from Maxwell's electromagnetic theory 343
153 Mass and angular momentum at the quasi-local level in general relativity 377
154 Objective state-vector reduction? 387
155 Embedding 2-surfaces in CM 393
156 A suggested further modification to the quasi-local formula 397
157 Higher-dimensional two-surface twistors 399
158 Numerical relativity by power series 403
159 Elements of twistor theory 409
160 Energy and its definition in general relativity 429
161 Two-surface twistors and hypersurface twistors 441
162 Exponentiating a relative H^2 445
163 Gravitational field mass 449
164 Gravity and quantum theory 467
165 Twistors and the time-irreversability of state-vector reduction 473
166 Gravity and state vector reduction 477
167 Big bangs, black holes and 'time's arrow' 497
168 Escher and the visual representation of mathematical ideas 525
169 Hermann Weyl, space-time and conformal geometry 541
170 Quantum physics and conscious thought 573
171 Local twistor transport at $\mathscr{I}^+$: An approach to the googly 601
172 An approach to a coordinate-free calculus at $\mathbb{I}$ 609
173 Twistors in general relativity 613
174 Newton, quantum theory and reality 633
175 On Qadir's intertwining operators 667
176 On the origins of twistor theory 671
177 New directions in quantum gravity 693
178 Pretzel twistor spaces 703
179 Twistors and state-vector reduction 713
180 Aspects of quasi-local angular momentum 717
181 Fundamental asymmetry in physical laws 727
182 Topological QFT and twistors: Holomorphic linking 741
183 Holomorphic linking: Postscript 745
184 Singularities and big-bang cosmology 747
185 On the physics and mathematics of thought 751
186 A twistor conformal field theory for four space-time dimensions 785
187 Minds, machines and mathematics 791
188 Tilings and quasi-crystals: A non-local growth problem? 811
189 Non-Hausdorff Riemann surfaces and complex dynamical systems 839
190 Using spinors in general relativity 845
191 A twistorial approach to the full vacuum equations 851

VOLUME 5

1990s

192 Difficulties with inflationary cosmology 1
193 Light rays near i^0: A new mass-positivity theorem 19
194 Extended Regge trajectories: 12 years after 27
195 Précis of *The Emperor's New Mind: Concerning computers, minds and the laws of physics* 29
196 Author's response – the nonalgorithmic mind 43
197 Twistor theory after 25 years – its physical status and prospects 59
198 Twistors, particles, strings and links 89
199 Mass positivity from focussing and the structure of space-like infinity 101
200 Twistor theory for vacuum space-time: A new approach 107
201 What does the big bang tell us about quantum gravity? 113
202 Twistors as spin 3/2 charges 123
203 Twistors as charges for spin 3/2 in vacuum 133
204 The mass of the classical vacuum 143
205 Twistors as spin 3/2 charges continued: SL(3,C) bundles 151
206 Response to Tony Dodd's 'Godel, Penrose and the possibility of AI' 159
207 The biggest enigma 161
208 $\mathcal{H}$-space and twistors 167
209 'Penrose and mathematical ability' 189

210 The most general {2,2} self-dual vacuum: A googly approach 193
211 On impulsive gravitational waves 199
212 Kinking and causality 205
213 Puzzle page 213
214 Solution to puzzle in TN34 215
215 Gravitational collapse of the wave function: New thoughts 221
216 Setting the scene: The claim and the issues 225
217 The modern physicist's view of nature 259
218 Gravity and quantum mechanics 311
219 Foreword to *Mr Tompkins in paperback* 323
220 Mathematics, physics, computers and conscious understanding 327
221 An example of indeterminacy in the time-development of "already unified field theory": A collision between electromagnetic plane waves 339
222 Orthogonality of general spin states 349
223 A simple observation concerning {2 2} vacuums 355
224 The orthographic image of a regular tetrahedron 357
225 Author's response – An Emperor still without mind 361
226 Square-root voting 369
227 On Bell non-locality without probabilities: More curious geometry 375
228 Quantum non-locality and complex reality 401
229 On the cohomology of impossible figures 415
230 Spin 3/2 fields and local twistors 419
231 Mathematical intelligence 427
232 Is conscious awareness consistent with space-time descriptions? 459
233 A twistor-topological approach to the Einstein equations 475
234 On the second law of thermodynamics 485
235 Twistors and the Einstein equations 491
236 What 'gaps'? Reply to Grush and Churchland 505
237 Non-locality and objectivity in quantum state reduction 521
238 Must mathematical physics be reductionist? 531
239 Concerning space-time points for spin 3/2 twistor spaces 547
240 In {2 2} vacuums, the null-datum on null rays is exactly r^{-5} 553
241 Why new physics is needed to understand the mind 557
242 The role of gravity in quantum state reduction 573
243 Consciousness involves noncomputable ingredients 593
244 Beyond the doubting of a shadow: A reply to commentaries on *Shadows of the Mind* 603
245 Chandrasekhar, black holes, and singularities 641
246 Conscious events as orchestrated space-time selections 661
247 Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness? 681
248 Space-time and cosmology 711
249 New ideas for the googly graviton 757
250 Incidence between complex null rays 771
251 Two puzzles 777
252 On gravity's role in quantum state reduction 779
253 Twistor theory, the Einstein equations, and quantum mechanics 801

VOLUME 6

1990s

254 Andrzej Mariusz Trautman 1
255 Twistor geometry of light rays 11
256 Remarks on tiling: Details of a $(1+\varepsilon+\varepsilon^2)$-aperiodic set 37
257 On understanding understanding 69
258 Mathematics: Pattern, precision and profundity 85
259 The mathematics of the electron's spin 109
260 Five things you can do with a surface in PT 115
261 Solutions to puzzles in TN41 121
262 Corrections to missprints in solutions to TN41 puzzles 125
263 Googly maps as forms 127
264 The need for a non-computational extension of quantum action in the brain 133
265 Chandrasekhar and the end of time 151
266 Science and the mind 161

267 Quantum state-reduction as a gravitational phenomenon; theory and experiment 165
268 Lionel Penrose: Colleague and father 169
269 Foreword to *Einstein's miraculous year: Five papers that changed the face of physics* 175
270 The question of cosmic censorship 185
271 Afterword to *The geometric universe: Science, geometry, and the work of Roger Penrose* 203
272 The twistor space for a strongly asymptotically flat vacuum 213
273 The complexity of our singular universe 223
274 Can a computer understand? 251
275 Spherically-symmetric solutions of the Schrödinger–Newton equations 279
276 Quantum computation, entanglement and state reduction 291
277 Foreword to *Ludwig Boltzmann: The man who trusted atoms* 305
278 Mathematics in the arts 309
279 Some remarks on twistor theory 323
280 The central programme of twistor theory 339
281 Lecture in honour of Leopold Infeld (extended outline only): Spinors in general relativity 371
282 On Ricci-flat twistor theory 381
283 Twistor theory and the Einstein vacuum 389
284 Introduction: Science at the turn of the millennium. 'Science into the next millennium: Young scientists give their visions of the future, part I.' 409

2000s

285 Introduction. *What is life?* 417
286 Drawing with complex numbers 433
287 Mathematical physics of the 20th and 21st centuries 441
288 Reminiscences of Christopher Strachey 459
289 On extracting the googly information 463
290 Wavefunction collapse as a real gravitational effect 489
291 On Bell non-locality without probabilities: Some curious geometry 509
292 The heritage of Pythagoras: Nineteen to the dozen 537
293 Twistors and general relativity 551
294 Towards a twistor description of general space-times: Introductory comments 579
295 Consciousness, the brain, and spacetime geometry: An addendum 597
296 The Schwarzschild singularity 605
297 The physics of the small and large: What is the bridge between them? 621
298 The rediscovery of gravity. The Einstein equation of general relativity 629
299 John Bell, state reduction, and quanglement 663
300 Consciousness, computation, and the Chinese room 677
301 Twistor geometry of conformal infinity 703
302 Gravitational collapse of the wavefunction: An experimentally testable proposal 713
303 Towards quantum superpositions of a mirror 719
304 On the instability of extra space dimensions 725
305 The problem of spacetime singularities: Implications for quantum gravity? 743
306 The remarkable role of mathematics in the physical universe 767

33

COSMOLOGICAL BOUNDARY CONDITIONS FOR ZERO REST-MASS FIELDS

The nature of time (ed. T. Gold), pp. 42–54. Cornell University Press, Ithaca, New York, 1968

The conference at which this lecture was given was held in Cornell University in the spring of 1963, or thereabouts, organized by Tommy Gold. The 'Mr X' who had a habit of asking awkward but insightful questions was Richard Feynman. It was on the transatlantic flight to attend this meeting that I obtained this geometrical way of looking at Rindler's notions of cosmological event horizon or particle horizon, where we see that these respective notions correspond to a spacelike nature for future conformal infinity (which arises when there is a positive cosmological constant), or for the conformally pictured big bang. See also Chapter 18.

2 Cosmological boundary conditions for zero rest-mass fields. *The nature of time* (ed. T. Gold), pp. 42–54

III. Cosmological Boundary Conditions for Zero Rest-Mass Fields

R. PENROSE

Suppose we consider solutions to Maxwell's equations in cosmological backgrounds. A question is: When can we call such a solution a purely retarded or purely advanced solution? We are helped in the discussion of these problems by the fact that all isotropic cosmological models are conformally flat, although this property is not essential to the discussion. For, by the conformal invariance of Maxwell's equations, we can, in such cases, reduce the problem to one concerning solutions of Maxwell's equations in flat space. The conformal-invariance property also holds for the Weyl neutrino field and, in an appropriate sense, for gravitation as well. We must be careful in our interpretations, however, because conformal transformations do not, in general, preserve the retarded or advanced nature of a field.

We can write the relevant source-free equations for spins $\frac{1}{2}$, 1, and 2 and for zero mass as follows:

$$\begin{array}{ll} \text{neutrino, spin } \tfrac{1}{2}: & \nabla^{A\dot{P}}\nu_A = 0, \\ \text{photon, spin } 1: & \nabla^{A\dot{P}}\phi_{AB} = 0, \\ \text{graviton, spin } 2: & \nabla^{A\dot{P}}\gamma_{ABCD} = 0, \end{array}$$

$$\text{where } \nabla^{A\dot{P}} = \sigma^{\mu A\dot{P}}\nabla_\mu.$$

We use two-component spinors throughout. Since the equations must hold in curved space, the differential operators are covariant derivatives. All these spinor quantities are symmetric in their indices. The spinor translation of the Weyl curvature tensor is γ_{ABCD}. If source fields are present, however, then we have terms on the right-hand sides of the equations. These equations have the property that they are conformally invariant. This is the key to the whole procedure. Conformal invariance means, of course, that the equations still hold if we make an "isotropic" space-time "stretch" at each point, which can vary from point to point. That is, using the same coordinates we can change the metric by a factor:

$$\hat{g}_{\mu\nu} = \Omega^2 g_{\mu\nu}; \qquad \hat{\sigma}_\mu{}^{A\dot{B}} = \Omega\sigma_\mu{}^{A\dot{B}},$$

where Ω is a function of position. The various "field equations" are invariant under these conformal transformations provided we take

$$\hat{\nu}_A = \Omega^{-3/2}\nu_A, \hat{\phi}_{AB} = \Omega^{-2}\phi_{AB}, \hat{\gamma}_{ABCD} = \Omega^{-3}\gamma_{ABCD}.$$

By means of such a transformation we can in most cases of interest turn infinity into an ordinary hypersurface, which is a boundary to the transformed space-time. We shall study fields in this transformed space-time instead of in the original physical one. We then have a more "finite" model, and we can, using *local* arguments, examine the behavior of the field at physical infinity. That is, if we examine components of the fields on the *boundary* of the transformed space, we can see what is meant by the radiation field in the *physical* space, and we can therefore define what is meant by purely advanced and purely retarded solutions. Now, the essential point is that in some cosmological models we get a spacelike surface at infinity, in some we get a null surface, and in others we get a timelike surface. Moreover, we may get different types of surface in the past and in the future. It turns out that the kind of surface we get is intimately related to the existence of "visual horizons" (also to the value of the cosmological constant), and to the behavior of radiation.

In all cases, we consider the infinity encountered in *null* directions. This is intimately related to the fact that it is the *zero* rest-mass field equations which are conformally invariant and such fields propagate in null directions. Space-times with null or spacelike surfaces at infinity turn out to be the most important. For Minkowski space, we in fact get a null cone as the future infinity and another null cone as the past infinity.

It must be emphasized that the ordinary null cones in space-time are not altered in the passage to the new conformally transformed space. This is an important feature of conformal invariance—that the null geodesics of the physical space are still null geodesics in the conformally transformed space. However, spacelike and timelike geodesics will not generally remain geodesics.

Now, let us consider, for example, a space with a spacelike boundary at future infinity. An observer's world line then intersects it at "time infinity." This intersection is a perfectly ordinary point on the boundary from the point of view of the transformed space. There will be at that point a past null cone, which divides the space into two regions, containing, respectively, events from which information can reach the observer and events forever hidden from him. Suppose we have, instead, a spacelike boundary surface in the past. Then the past null cones of spatially separated events near enough to this surface will not intersect (that is, such events will be "causally independent" of one another).

In asymptotically flat spaces, both infinites are null as in the case of Minkowski space. In the Einstein–de Sitter model, the future infinity is null. The past infinity is a singular point in the metric, but it so happens that we can do a conformal transformation so that the past infinity is no longer singular, and then it turns out to be spacelike. In the steady-state model, the situation is just the reverse: the future infinity is spacelike and the past infinity is null. This is, perhaps, not really quite fair, as the past infinity is not at infinity from the point of the view of the space-time metric, but it does not seem to make much difference whether infinity is "metrically" at infinity or not. You do not have to do any transformation in the past in the case of the steady-state model.

I must define the notions of "event horizon" and "particle horizon." Consider the totality of events which can be observed by a single observer (that is, consider the part of space-time traced out by the observer's past null cone). An "event horizon" occurs if this part is not the whole space, so that there are events which never become observable by the observer. If all events are ultimately observable by the observer, then there does not exist an event horizon. The event horizon divides the observable from the unobservable for a *given* observer (that is, taking the entire observer's world line as given). There are, of course, world lines of other observers which cross the given observer's event horizon. From the preceding discussion, it follows that *an event horizon exists when future infinity is spacelike* (Figure III-1). On the other hand, if future infinity is null (a null cone), the world lines of all observers (normally) pass through the vertex of this null cone when they reach the boundary. The ultimate null cones of all observers will then all coincide with the future boundary to the universe, so that *there is no event horizon when future infinity is null* (Figure III-2).

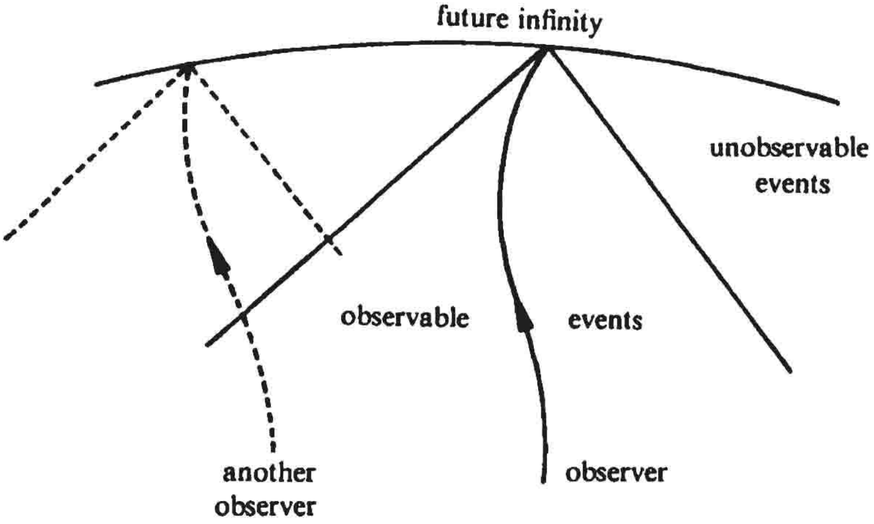

future infinity
unobservable events
observable
events
another observer
observer

FIGURE III-1

A "particle horizon" exists if, for an observer at an event in space-time, the "substratum particles" in the universe can be divided into two sets, those that he can see and those that he cannot see. This situation occurs when past infinity is spacelike. For, in this case, the starting points of the world lines of the substratum particles are spread out over the whole of past infinity. The past light cone of the observer (normally) intersects only a *subset* of these substratum particles, namely those that he can "see" at any one time. Thus, *a particle horizon exists when past infinity is spacelike* (Figure III-3). A characteristic feature of

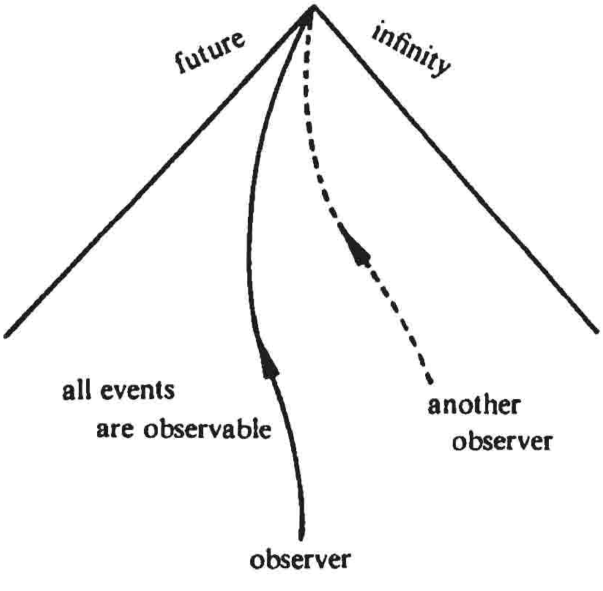

future
infinity
all events are observable
another observer
observer

FIGURE III-2

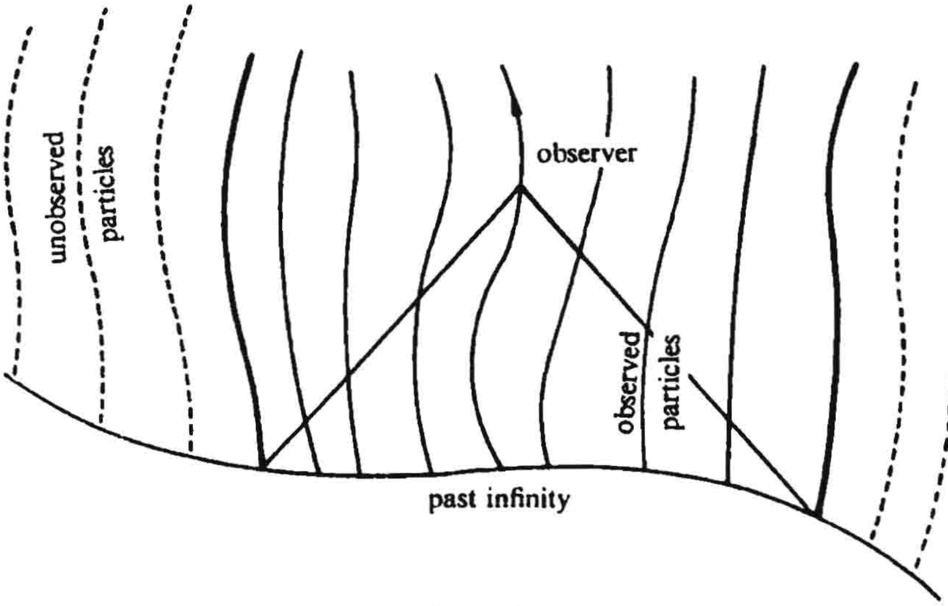

unobserved
particles
observer
observed
particles
past infinity

FIGURE III-3

particle horizons is that "waiting" helps; if the observer looks later he will see particles he could not see earlier. If past infinity is a null cone, then all the world lines of the substratum particles will originate at the vertex. The past light cone of any observer will then (normally) intersect *all* these world lines, so that *there is no particle horizon when past infinity is null* (Figure III-4). This is just the time reverse of the situation for the event horizon.

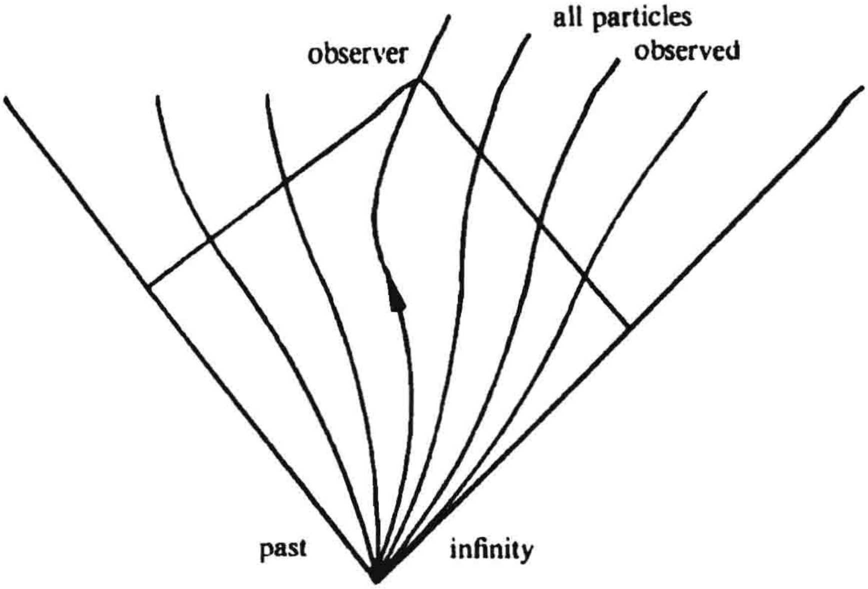

observer
all particles
observed
past
infinity

FIGURE III-4

WHEELER

Suppose that the inability of the observer to see certain events is due not to something in the geometry which intervenes between them, but to the fact that the observer crashes into a region of space having infinite curvature and gets killed. Is the concept of horizon valid in this case? I mention this on account of the similar features in the case of the Schwarzschild metric.

PENROSE

These horizons were defined originally for the Robertson-Walker metrics, and I think my discussion really applies only to models of this general kind. I'm not sure how one would extend the ideas to cover the other type of situations you refer to.

Now, it is possible to show quite generally that if there exists an event horizon (of the type I am considering) in a cosmological model, then purely advanced solutions of Maxwell's equations are not possible for arbitrary ("test") source distributions in that model. An example of such a model is the steady-state universe. We can also show generally that if a particle horizon exists in a cosmological model, then purely retarded solutions of the Maxwell equations are impossible for arbitrary sources. An example of this type is the Einstein–de Sitter model. The de Sitter hyperspherical space-time, which consists of two copies of the steady-state model, has both horizons. One "steady-state" branch goes into the past and the other goes into the future. If we stick them together, we get a complete hypersphere in five dimensions which has both types of horizon. In this case neither purely advanced solutions nor purely retarded solutions are possible for general source behavior.

There is more than one way of looking at the meaning of the purely advanced or retarded solution. The simplest way to visualize the situation for, say, a finite system of particles is to consider the behavior of the field as it proceeds along a null geodesic in space-time into the future or into the past. We can find a characteristic behavior, and we single out one term which behaves as $1/r$, where r is a suitably defined luminosity parameter along this geodesic. We can say that the radiation field is zero if the $1/r$ part vanishes. This procedure will work provided infinity is really at infinity, but if it is either singular or not "metrically" at infinity, we have a bit of trouble.

We examine the field along all null geodesics into past or future. Provided we have a proper infinity, we can ask whether the field falls off as $1/r$ or as $1/r^2$. If it always falls off as $1/r^2$ (or faster) into the past [future] then we can say that there is no incoming [outgoing] radiation. So a purely retarded [advanced] potential is one which has no $1/r$ term along the past [future] light cone. We shall consider a more general approach presently.

One of the big advantages of being able to bring infinity to a finite place is that we can do the initial-value problem for infinity in a perfectly rigorous way. We can specify data on the bounding surface. It is very important whether this surface is a null surface or a spacelike surface. If the surface is null, then we specify one complex number at each point of the surface. In the source-free case, this specification of one complex (that is, two real) numbers at each point of the past cone completely determines the solution. If we have a spacelike surface instead, we need four real numbers at each point, but they are usually rather hard to specify, even in the case of the electromagnetic field. We have in this case six numbers and two constraint equations. In the gravitational case things are still more complicated.

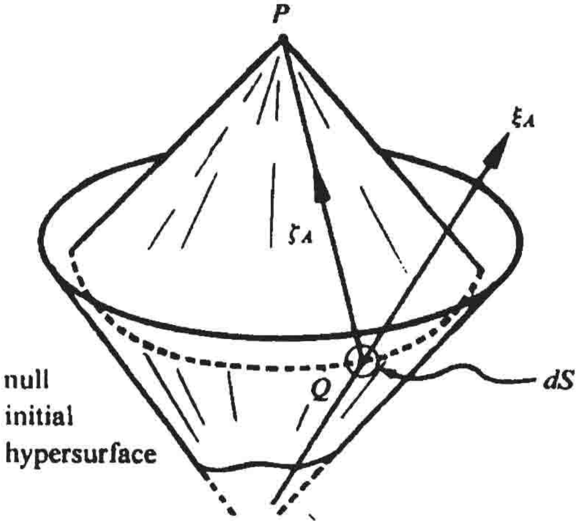


FIGURE III-5

In flat space, a zero rest-mass field can be described according to the following formula, which defines the field at a point P in terms of initial data. In Figure III-5, a *null* initial surface is represented. For arbitrary spin s, we have a spinor $\phi_{AB\ldots L}$ which is symmetric, having

$2s$ indices, satisfying $\nabla^{A\dot{P}}\phi_{AB\cdots L}=0$ if $s\geq\frac{1}{2}$, and satisfying $\Box^2\phi=0$ if $s=0$. The expression for $\phi_{AB\cdots L}$ at P, in terms of an integral involving data at the points Q which lie on the intersection of the null cone of P with the null initial surface, is

$$\phi_{AB\cdots L} = \frac{1}{2\pi}\int\frac{1}{r}\zeta_A\zeta_B\cdots\zeta_L\{D\phi - (2s+1)\rho\phi\}\,dS + \textit{any source terms.}$$

The quantities r, ζ_A, D, ρ, dS, and ϕ are defined by

$$(\overrightarrow{QP})_\mu = r\zeta_A\bar{\zeta}_{\dot{B}}\sigma_\mu{}^{A\dot{B}}; \qquad \xi^A\zeta_A = 1; \qquad D \equiv \xi^A\bar{\xi}^{\dot{B}}\nabla_{A\dot{B}},$$

the spinors ζ_A and ξ_A representing the two null directions orthogonal to dS, with dS = surface area element of 2-dim. intersection surface; $\rho=-\xi^B\nabla_{A\dot{B}}\xi^A$; and $D\xi^A=0$. The initial data quantity on a null initial hypersurface has one complex component:

$$\phi = \phi_{AB\cdots L}\xi^A\xi^B\cdots\xi^L.$$

This formula shows that the (free) field at the point P can be calculated by an integral over the intersection of a null initial surface with the past cone of P. The spinor associated with the null direction along the null cone of P is ζ_A and the quantity ϕ is the initial data quantity, which is just the field contracted with the spinor representing the null direction in the null initial surface. It is the projection of the field down into the initial surface, and it is just one complex number per point of the surface. Note that it is an integral just around an *intersection* of the null cone of P with the initial surface. This treatment works in flat space, and works with suitable modification of the factors in any conformally flat space, so it will work for any isotropic cosmology. If the space were not conformally flat, we would have to consider scattering effects. But since here it is only the data on the intersection of the null cone of P with the initial surface that comes in, we can apply the formula equally well when we have a spacelike initial surface. We can always find a null surface which coincides with a spacelike surface at its intersection with the null cone of P. Thus, if we can find ϕ in terms of the data on the spacelike surface, we can still apply the formula to find the field at P. The spinor ξ_A used here to define ϕ simply corresponds to the null direction (other than that of ζ_A) which is orthogonal to the intersection surface.

Let us now apply the formula when the initial surface is the surface representing past infinity. Then, the initial data quantity ϕ represents what, to an observer at *P*, would appear to him to be the free incoming radiation. If there is no incoming radiation, the $\phi_{AB\ldots L}$ is given entirely by the source terms. If the free field integral is zero for all points *P*, then the data must be zero. This means that there is no contribution to the field from the arbitrarily specifiable data at infinity. If we have a charge whose world line cuts the null cone of *P* at a specified point, then the contribution would come only from that point. More generally, we might have many sources and incoming field. Then we have a contribution from each intersection point of the world line of a charge with the null cone of *P* ("retarded contributions") together with the free field integral over initial data for the field at infinity ("incoming field").

The idea is that we try to specify initial data at infinity so that, for a retarded field, the free field integral is always zero. This amounts to the quantity ϕ being zero all over the initial surface (and for all allowable choices of ξ^A). When the initial surface is spacelike, this implies that $\phi_{AB\ldots L}$ is zero throughout the initial surface. This is generally impossible if $s>\frac{1}{2}$, since too many conditions (>4) on the initial data must be satisfied.

X

How do we calculate the field contribution at *P* from the source? The question is which way we solve the equations, because there is nothing unique about the source terms. We can use a different sort of analysis in the calculation; we can change the contributions which we consider to come from the sources and those from the surface. Perhaps the source terms can be changed by a homogeneous solution. We might say that this is equivalent to changing the incoming condition.

PENROSE

Yes, it is true that I haven't specified how to treat the source terms here, but this doesn't really affect the argument. If we consider, for simplicity, a single charge, in a space with a spacelike past infinity, and work out the field at a point *P* whose past cone does not even intercept the charge (that is, the charge is outside *P*'s particle horizon), we have the situation illustrated in Figure III-6. This single charge is the only source here, and since the past cone of *P* does not meet the world line of the charge, we would have to say that the entire field at *P* is due to incoming radiation field. Thus, for a purely retarded field, the field at *P* would have to be *zero!* However, it is impossible to arrange that the field be zero at *all* such points *P*, as we can see by examining the field on the cone (shown dotted in the diagram) which is the future cone of the initial point (at infinity) of the world line of the charge. It turns out that there must be a discontinuity in the field along this cone and that the free-field equations are violated at this discontinuity.

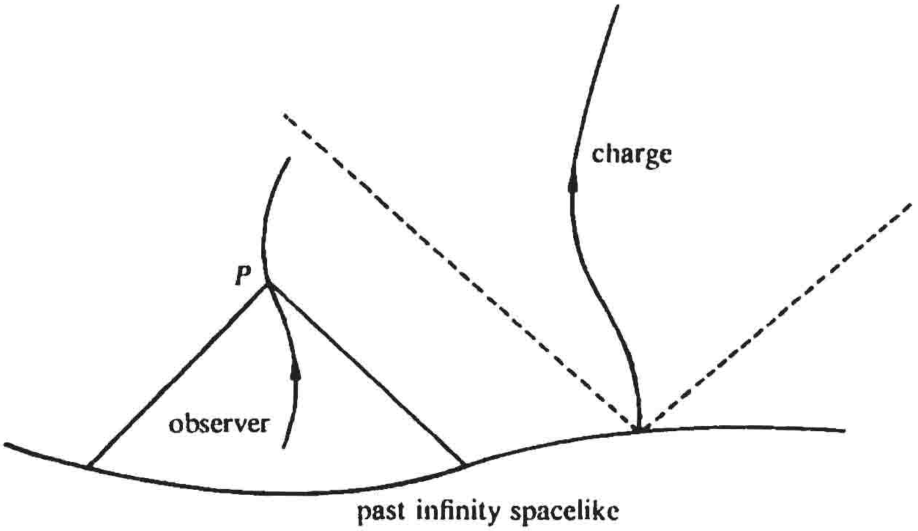


FIGURE III-6

SCHIFF

You could use the fact that the charge can be measured as an integral of field quantities over a sphere surrounding the world line to show this violation.

PENROSE

That's a very good point. A two-sphere surrounding the charge but lying outside the dotted cone would lie entirely in the *zero* field region. The integral giving the charge surrounded would thus be zero. This is a contradiction, so it follows that a purely retarded field is impossible for this situation! This can also be seen if we go back to the original space. Then the world line of the particle extends down forever; it always has a field and its field cannot suddenly cut off at the particle horizon.

BONDI

How do we know that the field at P is all radiation field? I see that the distinction between radiation and Coulomb fields disappears under conformal transformations, since we cannot distinguish any more the $1/r$ from the $1/r^2$ part. But how then do we make the distinction appear again? Cannot the particle move in such a way that it produces only a Coulomb field?

PENROSE

The definition of radiation field is the contribution to the integral which does not come from sources. There are no sources outside the dotted cone, so that the complete contribution is the radiation field outside. The condition that there is no incoming radiation therefore implies that the field must be zero everywhere outside the dotted cone. What looks like a Coulomb field in this picture is radiation in the untransformed field. The conformal transformation does not preserve the concept of radiation field, but the radiation field can be easily identified in the transformed case in terms of field components on the hypersurface representing infinity.

MORRISON

Are you not placing a special regularity condition, a special condition at infinity, on Maxwell's equations, which we normally do not impose?

PENROSE

In a sense, what I do is equivalent to imposing a regularity condition at infinity, but it apparently turns out that any field which looks reasonable in the physical space, and which falls off in the right way, corresponds to a field which is finite on the initial surface, and the formalism can therefore be applied.

HOGARTH

Can you not change a spacelike infinity into a null infinity?

PENROSE

No, because conformal invariance leaves null lines invariant.

HOYLE

I see that the spacelike or null character of the surface affects in a very intimate way the choice of data on the surface. Is this something connected with the general behavior of partial differential equations in the choice of initial data on null cones?

PENROSE

Yes, but this problem is something I am not completely clear about myself.

BONDI

Surely the point is that we can have a boundary between zero and nonzero things only on a null cone. This is a very crucial place, the only place where nonanalytic behavior can take place. This is where the propagation of discontinuities occurs.

PENROSE

This business about having two pieces of data when the suface is null and four when it is spacelike is something that I do not fully understand. There can be a surface which goes along spacelike, then turns and becomes null, then becomes spacelike again. We must then specify four pieces of data first, then two, then four again, as in Figure III-7.

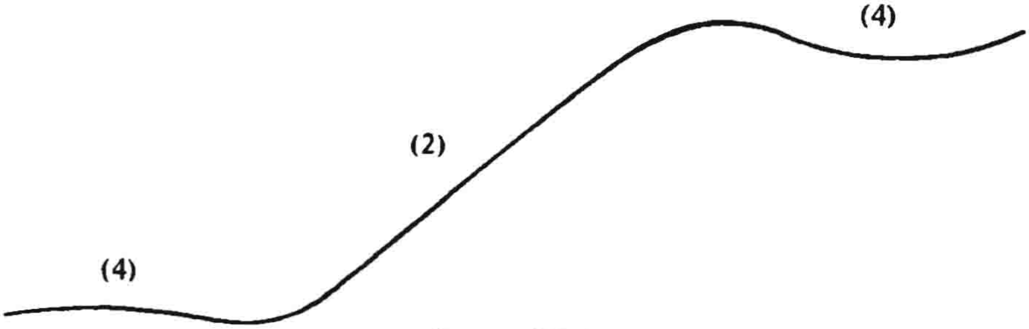


FIGURE III-7

BONDI

The surface gets only the news that is broadcast to a point on it. When we are, say, in the null piece, we travel with one stream of news, and therefore we intercept only one other stream of news. We must think of the information as sliding along with us. We are swimming with the waves. This is quite reasonable.

My problem is more a psychological than a physical one, but perhaps yours is the real explanation.

SCIAMA

This analysis applies only for a finite number of particles. Do you suppose this treatment could be extended to cover the case of an infinite number of particles?

PENROSE

I feel sure that this type of approach would be valuable, but the detailed physical questions soon begin to arise; these worry me. I am trying only to construct conditions which make mathematical sense to me. The situation might be changed if the universe is treated as full of particles; I do not know. My suggestions hold if there are just a few particles with a cosmological background.

34

ON GRAVITATIONAL COLLAPSE AND COSMOLOGY

Gravity research foundation essay prize, 1968. With S. W. Hawking

Stephen Hawking and I won the Gravity Research Foundation's Second Prize with this essay (compare Chapter 27). The full write-up is Chapter 42.

On Gravitational Collapse and Cosmology

by

Stephen Hawking

Department of Applied Mathematics and Theoretical Physics,
University of Cambridge.

and

Roger Penrose

Department of Mathematics, Birkbeck College,
University of London.

Abstract

We present a new theorem on space-time singularities. On the basis of the Einstein (or Brans-Dicke)theory, and without using any Cauchy surface assumption, we show (essentially from the property that gravitation is always attractive) that singularities will occur if there exists *either* a compact spacelike hypersurface *or* a closed trapped surface *or* a point whose past light-cone starts converging again. The first condition would be satisifed by any spatially closed universe, the second by a collapsing star and the third by the observable portion of our actual universe - as we shall show follows from observations of the microwave background radiation.

1. Introduction

The most immediately noticable fact about gravitation is that it always seems to be attractive. Thus, despite its extreme weakness compared to electromagnetism, it is nevertheless significant for a large body such as a star since all the gravitational fields of individual particles add up while the electromagnetic fields cancel each other out. Indeed on an even larger scale, that of the whole Universe, gravitation dominates over all other forces. This attractive character and its r^2 dependence make gravitation rather a 'dangerous' force since it is potentially unstable: if a star is compressed slightly, the particles move closer together and so the attractice force between them will increase. Normally, of course, the repulsive pressure forces would increase by a somewhat greater amount and so restore balance. However, Chandrasekhar[1] has shown that when a star of greater than 1.5 times the solar mass exhausts its nuclear fuel and cools, the pressure forces are insufficient to resist the gravitational attraction. What then would happen to such a star? Would it collapse to some sort of singularity or would it happen that the smallest departure from spherical symmetry would cause different parts not to fall exactly towards the centre and so miss each other.

In the reverse direction in time a similar question arises in respect of the whole Universe: Was there a singularity in the past or did the Universe manage somehow to pass smoothly from a contracting phase to the present expansion? In this essay we shall present a new and very general theorem which shows that singularities would be expected both in the collapse of a star and at the beginning of the expansion of the Universe. This theorem combines and extends two previous theorems of the authors[2,3]. It is based on general relativity as this the most satisfactory theory of gravitation so far proposed but similar results would probably hold in any relativistic theory in which gravitation is attractive. In particular the theorem applies also to the theory of Brans and Dicke[4] so long as gravitation remains attractive. (If the gravitational constant were to change sign, as is in principle allowed in the Brans-Dicke theory, this could only occur via a region where it becomes infinite. In such a region, the presence of any matter would cause infinite curvature).

2. The Theorem

One might define a space-time singularity as a point at which the metric is degenerate or at which the curvature is infinite. However, such a point ought not to be regarded as part of space-time since the known laws of physics could not hold there. We shall therefore adopt the view that space-time consists only of points at which the metric is Lorentzian and suitably differentiable (say C^2). We then detect singular points which have been "cut out" by the existence of incomplete geodesics. This approach is discussed further in references 5 and 6.

In the first theorem[2,7,8] to use this criterion for a singularity, it was shown that there must exist incomplete null geodesics if

1) The energy-momentum tensor obeys the inequality*

$$T_{ab} V^a V^b \geqslant \tfrac{1}{2} V_a V^a T$$

for any timelike vector V^a.

2) There exists a noncompact spacelike hypersurface H which is a Cauchy surface (i.e. it intersects every inextendible timelike or null curve once and once only).

3) There exists a closed trapped surface T(i.e. T is a compact spacelike 2-surface such that both the 'ingoing' and the 'outgoing' families of null geodesics orthogonal to T are converging: see Fig.1).

Condition 1) is a very reasonable inequality which is satisfied by all known forms of matter. It is the condition that gravity should always be attractive. Condition 3) will be satisfied when a star~~s~~ collapses inside its Schwarzschild radius. One would expect stars of greater than 1.5 - 15 solar masses to do this eventually provided that their angular momemtum is sufficiently small. (This uncertainty in the mass limit results from uncertainty as to how much material is ejected during the collapse process.

The theorem will also hold if condition 3) is replaced by 3'):

3') There is a point p such that the divergence of the system of null geodesics through p changes sign somewhere to the part of p (Fig.2).

---

* In refs. 2,7,8 the effectively weaker assumption $R_{ab} l^a l^b \leqslant 0$, for null vectors l^a, was all that was used. This has the advantage that the presence of a cosmological constant does not affect the discussion. However, it is hard to believe that a non-zero λ can produce any qualatively noticeable affects in regions of high curvature. Here we assume $\lambda = 0$. (Note that $\lambda > 0$ implies that gravitation is not 'always attractive').

It is shown in the Appendix that observations of the cosmic background radiation indicate that there is indeed enough matter on our past light-cone to cause it converge again.

It this seems reasonable to assume that conditions 1) and 3) or 3′) are satisfied. However the weakness of the theorem lies in condition 2) which requires there to exist a noncompact Cauchy surface. That it should be non-compact is not much of a restriction since it was shown in reference 3 that there will be incomplete geodesics in any 'generic' solution which satisfied condition 1) and which had a compact Cauchy surface~~s~~. However the requirement that there should exist any global Cauchy surface at all is a very strong one. For being a Cauchy surface is a property not of the surface itself but of the whole space-time in which it is imbedded. There are plenty of solutions known which do not have Cauchy surfaces. Thus one might feel that this theorem does not indicate that we must expect singularities in space-time, but merely that global Cauchy surfaces do no exist in our Universe. To remove this weakness we shall present a new version of the theorem ~~t~~which does not require existence of a Cauchy surface This new theorem also~~w~~ shows that there will be a singularity if the Universe is spatially closed, that is, if there exists a compact spacelike surface (not necessarily a Cauchy surface).

The precise statement of the theorem is : Space=time is not timelike and null geodesically complete if,

(a) condition 1) holds.

(b) Every timelike and null geodecis has a point on it at which

$k_{[a}R_{b]cd[e}k_{f]}k^c k^d$ is nonzero, where k^a is the tangent vector to the geodesic.

(c) There are no closed timelike curves.

(d) One of the three following conditions holds:

(i) condition 3).

(ii) condition 3′).

(iii) There is a compact spacelike hypersurface H.

* With condition (b) this implies strong causality[11].

Condition (b) would be expected to hold in any 'generic' space-time. (It fails in certain highly spacial exact solutions, however, but this is not of interest physically). In the presence of (a), condition (b) will follow if every timelike geodesic encounters some curvature and every null geodesic contains a point at which it fails to be directed along a principal null direction of the Weyl tensor. In a physically reasonable solution, the presence of randomly oriented gravitational field (e.g. radiation) is to be expected. This would be sufficient to ensure that (b) holds.

3. Proof

We shall give the proof for cases (ii) and (iii) that for (i) is similar to (ii).

For a set S we let $I^+(S)$ and $J^+(S)$ denote the sets of points which can be reached from S by future directed timelike and nonspacelike curves respectively. We define $E^+(S)$ to be $J^+(S) - I^+(S)$. Points of $E^+(S)$ will lie on the boundary of $J^+(S)$ and will be reached from S by null geodesics. All these definitions have duals in which future is replaced by past and '+' by '-'. We define $C^+(S)$ as the set of points $p \in J^+(S)$ such that $J^-(p) \cap J^+(S)$ is a non-empty compact set on which strong causality held. We denote by $D^+(S)$ the set of all points q such that every inextensible past directed nonspacelike curve through q intersects S.

We shall assume that space-time is timelike and null geodesically complete and show that this leads to a contradiction with conditions (a), (b), (c) and (d). Consider first case (ii). $\dot{J}^-(p)$, the boundary of $J^-(p)$ will be generated by null geodesic segments which may have a past end-point but which can have a future end-point only if they pass through p. Near p, $\dot{J}^-(p)$ will be generated by the past-directed null geodesics through p. As condition (a) holds and the past directed null geodesics from p start converging again, there would be a point conjugate to p on every such geodesic within a finite affine distance from p. Points on such a geodesic beyond the conjugate point could be connected to p by a timelike curve and so would lie in the interior of $J^-(p)$. Thus $E^-(p)$ would be compact.

Let A denote $E^-(p)$ in case (ii) and H in case (iii). Then in both cases $A = E^-(A)$ will be compact. This implies that $\dot{J}^-(A) - A$, if nonempty, will be generated by null geodesic segments which have no future end-point . Now consider $C^-(A)$ which will contain $D^-(A)$[5]. Suppose $q \in C^-(A)-D^-(A)$ exists. Then there would be an inextendible future directed curve λ from q , not meeting A. Since $J^+(q) \cap J^-(A)$ would be compact and strong causality holds, λ would have to intersect $J^-(A)-A$ at some point r. But there would be an inextendible future directed with geodesic in $\dot{J}^-(A)$ through r. This would mean that $J^+(q) \cap J^-(A)$ was not compact. This shows that $C^-(A)$ equals $D^-(A)$.

Figure 3 shows possible forms for $C^-(A)$ in the two cases. However, it should be emphasized that the diagrams are meant for illustration only and that there are other possibilities.

By an argument similar to one in reference 5, $C^-(A)$ cannot have compact closure. For if it had, we could cover $C^-(A)$ with a finite number of local causality neighbourhoods B_i. Then if $p_1 \in J^-(A) \cap [B_1 - \bar{C}^-(A)]$ there would be a future directed nonspacelike curve λ_1 , from p_1 to A which intersected $\dot{C}^-(A)$ in some other neighbourhood B_2. Continuing this process would exhaust all the B_i and so lead to a contradiction. There must be an inextendible past directed nonspacelike curve λ from the compact set A which remains in $C^-(A)$ since if every such curve left $C^-(A)$ it would have compact closure. As $E^-(A) - A$ is empty in both cases, it is possible to choose λ to be timelike. The boundary of $J^+(\lambda)$ will be generated by null geodesic segments which cannot have past end-points. By condition (b), each of these segments will therefore have to have a future end-point as otherwise there would be points of $J^+(\lambda)$ which had timelike separation. Every inextendible future directed nonspacelike curve from λ will intersect A. Thus if F denotes the compact set $A \cap \bar{J}^+(\lambda)$ then $E^+(F) - F$ will consist of those points of $\dot{J}^+(\lambda)$ through which there are null geodesic generating segments which intersect A to the past. Since each such segment must have a future end-point, it follows that the set $G = F \cup E^+(F)$ is compact,

We then use similar arguments to those above to show that $\bar{D}^+(G)$ is noncompact and that there is an inextensible timelike curve μ which remains in $D^+(G)$. There will be points $q \in \lambda$ and $r \in \mu$ such $r \in I^+(q)$. Thus there will be a curve α which is inextensible in both future and past directions and which remains in $D^-(F) \cup D^+(G)$. Let a_i and b_i be sequences of points on $\alpha \cap I^+(F)$ and $\alpha \cap I^-(F)$ respectively such that any compact segment of α contains a finite number of each sequence. For each value of i there will be a timelike geodesic segment γ_i of maximum length between a_i and b_i[5]. Each γ_i will intersect the compact set F. Thus there will be a $q \in F$ which is a limit of $\gamma_i \cap F$ and a nonspacelike direction at q which is a limit of the directions of the γ_i. Let γ be the inextendible geodesic through q with this direction. It will remain in $D^-(F) \cup D^+(G)$ both in the future and in the past. By condition (b), there will be conjugate points x and y of γ. As the positions of conjugate points on a geodesic are determined by an integral of the curvature along the geodesic, they can be chosen as a continuous function of the position of the intersection of the geodesic with F and of the direction at F. Thus if U and V are any neighbourhood of x and y respectively, there will be some γ_i which intersects U and V and which contains conjugate points x' and y'. But this is impossible[9] as γ_i is supposed to be a geodesic segment of maximum length between a_i and b_i. This establishes the desired contradiction which shows that the original assumption that space-time is timelike and null geodesically complete must be false.

Appendix

We wish to show that there is sufficient matter on our past light-cone to cause it to start converging again. A sufficient condition for this to be so is that there should be a distance R such that along every past directed null geodesic from us,

$$\frac{R}{3}\int_{R}^{\infty}\frac{8\pi G}{C^2}\,T_{ab}K^{a}K^{b}\,dr > 1 \qquad (I)$$

In this integral, $K^a = \frac{dx^a}{dr}$ is the tangent vector to the null geodesic and r is an affine parameter on the null geodesic normalised so that at is, $r = 0$ and $K^a U_a = \frac{1}{C}$ where U^a is the past directed unit timelike vector representing the local standard of rest.

In a forthcoming paper[10], it is shown that, with certain assumptions, observations of the microwave background radiation indicate that not only do the past directed null geodesics from us start converging again but so also do the timelike ones, As we are concerned only with the null geodesics, the assumptions we shall need will be weaker.

The observations show that between the wavelengths of 20 cm and 2mm the background radiation is isotropic to within $1^{o}/o$ and has the spectrum of a black body at 2.7°K. We shall assume that this black body spectrum indicates not that the radiation was necessarily created with this form, but that it has been thermalised by repeated scattering. Thus there much be sufficient matter on each past directed null geodesic from us to make the optical depth large in that direction. We shall show that this matter will be sufficient to cause the inequality I to be satisfied.

The smallest ratio of density to opacity at these wavelengths will be obtained if the matter consists of ionised hydrogen in which case there would be scattering by free electrons. The optical depth would be

$$\int_0^\infty \frac{\sigma}{m} \rho \, CK^a U_a \, dr \qquad \text{(II)}$$

where σ is the Thomson scattering cross-section, m is the mass of a hydrogen atom, ρ is the density of the ionised gas and U^a is the local velocity of the gas. The red-shift Z of the gas is given by $(CK^aU^a - 1)$. We assume that this increases down out past-light cone. As galaxies are observed with red-shifts of 0.46 mostof the scattering must occur at red-shifts greater than this. (In fact if the quasars really are at cosmological distances, the scattering must occur at red-shifts of greater than 2). With a Hubble constant of 100 km./ sec./ Megaparsec, a red-shift of 0.46 corresponds to a distance of about 3×10^{27} cms. Taking R to be this distance, the contribution of the gas density to the integral in I is

$$1.8 \int_R^\infty \rho \, C^2 (K^a U_a)^2 \, dr$$

while the optical depth of gas at red-shifts greater than 0.46 is

$$4 \int_R^\infty \rho \, C \, K^a \, U_a \, dr$$

As $C K^a U_a$ will be greater than 1.46 for r greater than R, it can be seen that the inequality I will be satisfied at an optical depth of about 1.6. If the optical depth of the Universe were less than this one would not expect a black body spectrum as the photons would not suffer sufficient collisions to thermalise them. Even if the radiation were to be created with a black body spectrum, what one would see would be a dilute 'grey' body spectrum which could agree with the observations between 20 cms. and 2 cms but which would not fit those at 9 mm and 2 mm. Thus we can be fairly certain that condition 3') is satisfied in the observed Universe.

References

1. Chandrasekhar, S. Mon Not. R.A.S. 95 207 (1935).
2. Penrose, R. Phys.Rev.Lett. 14 57 (1965).
3. Hawking, S.W. Proc.Roy.Soc. A, 295 490 (1966).
4. Brans, C. amd Dicke, R.H. Phys.Rev. 124 3 (1961).
5. Hawking, S.W. Proc. Roy.Soc. A. 300, 187 (1967).
6. Geroch, R.P. 'What is a singularity in general relativity' (Preprint).
7. Penrose, R. Lectures delivered at the 1967 Battelle Summer Rencontres. To be published by Benjamin Inc. New York.
8. Penrose, R. Adams Prize Essay 1966 University of Cambridge.
9. Hawking, S.W. Proc.Roy.Soc. A 294 511 (1966).
10. Hawking, S.W. and Ellis, G.F.R. Ap.J. to appear.
11. Hawking, S.W., Adams Prize Essay 1966 University of Cambridge.

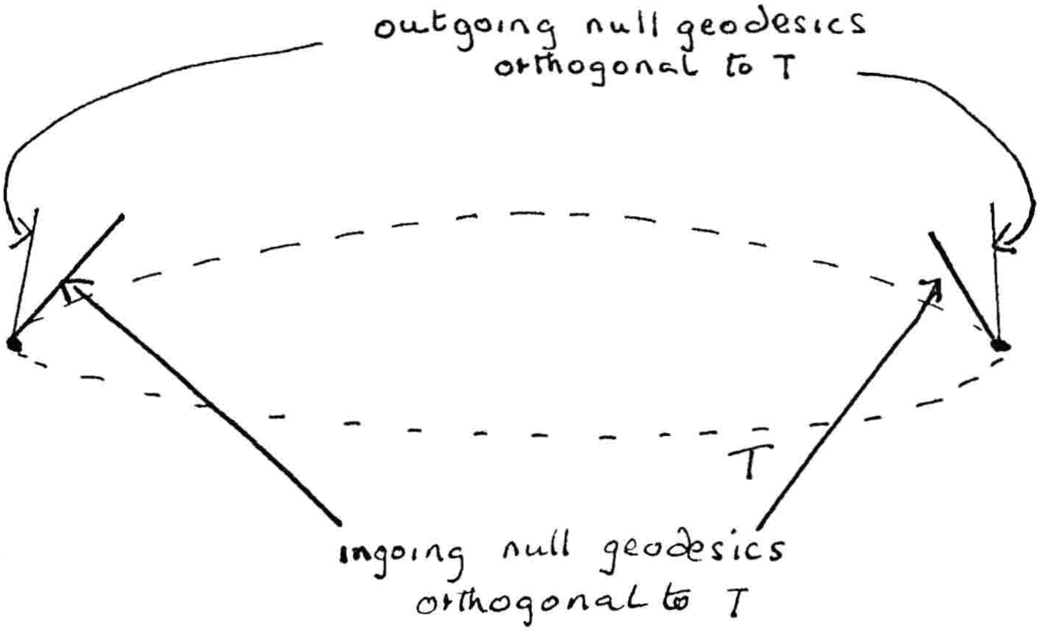

outgoing null geodesics orthogonal to T
T
ingoing null geodesics orthogonal to T

FIGURE I

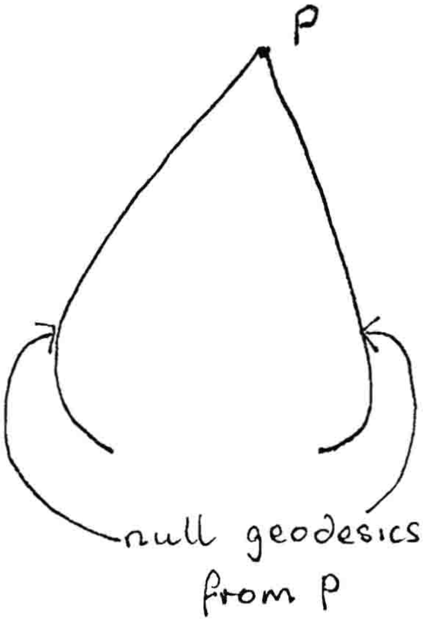
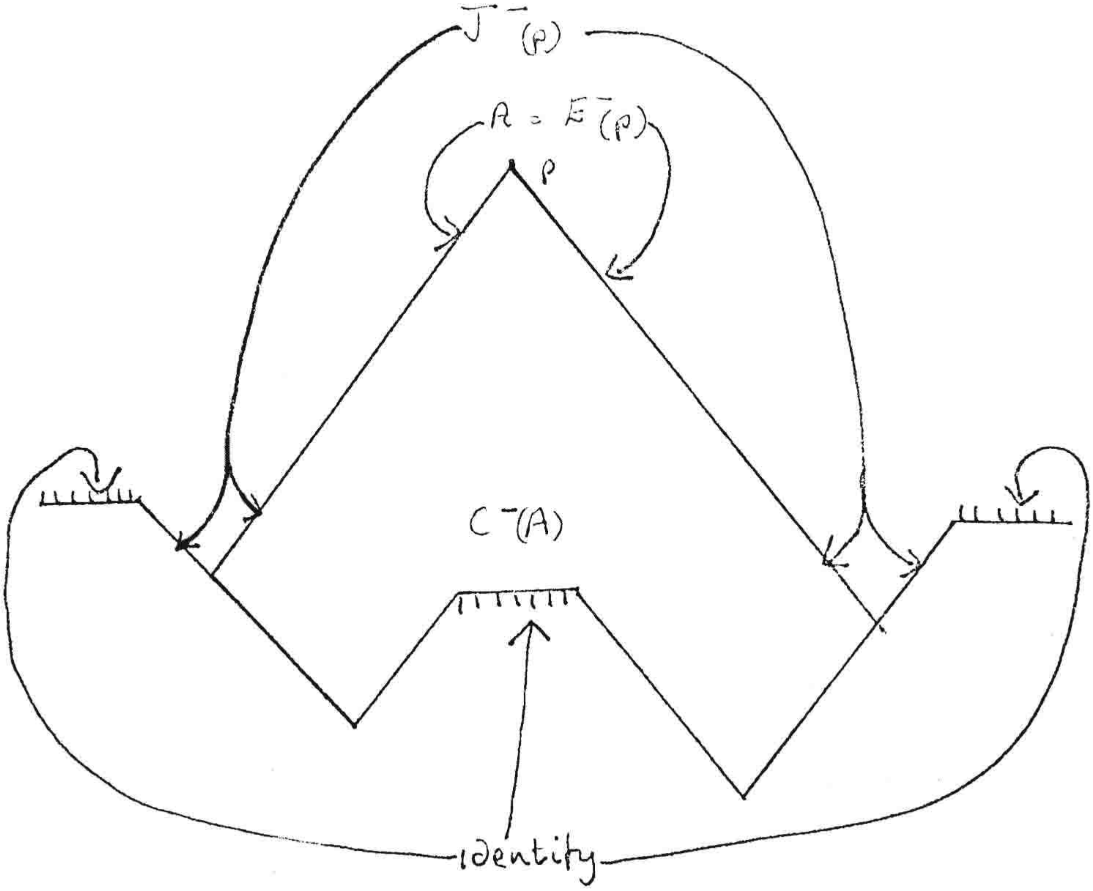

p
null geodesics from p

FIGURE 2.

case (ii)


$J^-(p)$
$A = E^-(p)$
p
$C^-(A)$
identity

case (iii)

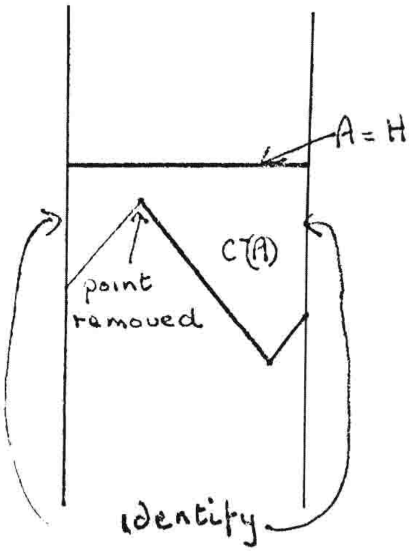

A = H
$C^-(A)$
point removed
identity

FIGURE 3.

I was born in 1942 and was educated at St. Albans School and University College, Oxford, where I took a B.A. in Physics. I then went to Cambridge to do research in Cosmology and General Relativity under the supervision of Dr. D. W. Sciama. I obtained my Ph.D. in 1966 and in the same year won an Adams Prize. Since 1965 I have been a Research Fellow of Gonville and Caius College, Cambridge.

I am married with one son.

Stephen Hawking

ROGER PENROSE

Born: 8th August 1931, Colchester, Essex, England.

Degrees: B.Sc. (Spec.) University College, London (1st class) 1952
Ph.D. Cambridge 1957

Employment, etc.:

Feb. 1956 - Aug. 1956: National Research Development Corporation, Tilney Street, London. (Mathematical work in connection with electronic computer programmes). Also consultant 1956-57.

1956-1957 Assistant Lecturer (Pure Mathematics) Bedford College, London.

1957-1960 Research Fellow, St. John's College, Cambridge.

1959-1961 N.A.T.O. Research Fellow, at Princeton, N. J., U.S.A. and also at Syracuse, N. Y. and Cornell, Ithaca, N.Y. (General Relativity and Quantum Theory)

1961-1963 Research Associate under U.S.A.F. contract, King's College, London (Mathematics Department)

1963-1964 Visiting Associate Professor (Mathematics and Physics), University of Texas, Austin 12, Texas, U.S.A.

1964-1966 Reader; 1966-Professor (Applied Mathematics), Birkbeck College, London.

During academic year 1966-1967, on leave in U.S.A. at Yeshiva University, New York; Princeton University; Cornell University; University of Chicago; Battelle Institute Seattle)

35

NEW CONSERVATION LAWS FOR ZERO REST-MASS FIELDS IN ASYMPTOTICALLY FLAT SPACE-TIME

Proc. Roy. Soc. Lond. Ser. A **305(1481)** (1968), 175–204. With E. T. Newman

This is the full write-up of the work of Chapter 23. As an incidental point, it may be remarked that the divergence-free energy-momentum tensor for the conformally invariant scalar wave equation that we wrote down in the second footnote on p. [194] is the same as that later discovered by Callan, C.G., Coleman, S., and Jackiw, R. (1970) A new improved energy-momentum tensor, *Ann. Phys.* **59**, 42–73.

30 New conservation laws for zero rest-mass fields in asymptotically flat space-time. *Proc. Roy. Soc. Lond. Ser. A* **305(1481)** (1968), 175–204

Proc. Roy. Soc. A. **305**, 175–204 (1968)
Printed in Great Britain

New conservation laws for zero rest-mass fields in asymptotically flat space-time

By E. T. Newman
University of Pittsburgh, Pittsburgh Pa., U.S.A.

and R. Penrose†
Birkbeck College, London, England

(*Communicated by H. Bondi, F.R.S.—Received* 14 *June* 1967—
Revised 15 *December* 1967)

Some recently discovered exact conservation laws for asymptotically flat gravitational fields are discussed in detail. The analogous conservation laws for zero rest-mass fields of arbitrary spin $s(=0, \frac{1}{2}, 1, \ldots)$ in flat or asymptotically flat space-time are also considered and their connexion with a generalization of Kirchoff's integral is pointed out.

In flat space-time, an infinite hierarchy of such conservation laws exists for each spin value, but these have a somewhat trivial interpretation, describing the asymptotic incoming field (in fact giving the coefficients of a power series expansion of the incoming field). The Maxwell and linearized Einstein theories are analysed here particularly. In asymptotically flat space-time, only the first set of quantities of the hierarchy remain absolutely conserved. These are $4s+2$ real quantities, for spin s, giving a $D(s, 0)$ representation of the Bondi–Metzner–Sachs group. But even for these quantities the simple interpretation in terms of incoming waves no longer holds good: it emerges from a study of the stationary gravitational fields that a contribution to the quantities involving the gravitational multipole structure of the field must also be present. Only the vacuum Einstein theory is analysed in this connexion here, the corresponding discussions of the Einstein–Maxwell theory (by Exton and the authors) and the Einstein–Maxwell–neutrino theory (by Exton) being given elsewhere. (A discussion of fields of higher spin in curved space-time along these lines would encounter the familiar difficulties first pointed out by Buchdahl.)

One consequence of the discussion given here is that a stationary asymptotically flat gravitational field cannot become radiative and then stationary again after a finite time, except possibly if a certain (origin independent) quadratic combination of multipole moments returns to its original value. This indicates the existence of 'tails' to the outgoing waves (or back-scattered field), which destroys the stationary nature of the final field.

1. Introduction

Some new conserved quantities, associated with the asymptotic gravitational field in Einstein's (vacuum) theory have been briefly described elsewhere (Newman & Penrose 1965; see also Penrose 1967).‡ For the electromagnetic and neutrino fields in the Einstein theory, additional conserved quantities also analogously exist. The purpose of the present paper is to discuss and analyse these various quantities in greater detail than has been given hitherto and to furnish proofs of some statements made in the first publication.

It is well known that for an isolated system in ordinary electromagnetic theory the total charge is absolutely conserved, while in the *linearized* gravitational theory (cf. Sachs & Bergmann 1958) the total mass, linear and angular momentum are correspondingly absolutely conserved. That is to say, these quantities are conserved in the strong sense that there is no contribution carried away by outgoing waves of the field itself; the quantities retain their original values even after the emission of radiation. (The conservation, in each case, depends only on the field equations holding in the asymptotic region.) On the other hand, in the full non-linear gravitational theory, the mass and momentum are no longer absolutely conserved by *can* be carried away by the radiation field, so as to give a positive energy flux at infinity (cf. Bonnor 1959; Bondi 1960; Bondi, van der Burg & Metzner 1962; Sachs 1962*a*; Newman & Unti 1962; Møller 1963; Penrose 1964; Cornish 1965; Bonnor & Rotenberg 1966; Winicour 1967). It had consequently been commonly believed that no absolutely conserved quantities, in this sense, exist in the Einstein gravitational (vacuum) theory and that only the total charge is absolutely conserved in the Einstein–Maxwell theory. It was therefore a considerable surprise to discover that, in fact, there are ten quite different quantities (defined in (4·19); cf. also (3·31) for the corresponding linear theory quantities) which *are* absolutely conserved for asymptotically flat Einstein vacuum fields, with six more (defined in (5·18); cf. (3·17)) for Einstein–Maxwell fields and four additional quantities (cf. (5·18)) if Weyl neutrino field† is also present. The original ten quantities of the pure gravitational theory are still conserved in the coupled theories, but only if they are modified by the addition of extra terms arising from the Maxwell and Weyl neutrino fields. (These modifications are discussed elsewhere; Exton, Newman & Penrose 1968; Exton 1967). No such modification is necessary in the case of the Maxwell or Weyl neutrino conserved quantities.

---

† On leave, for the year 1966/67, at Yeshiva, Princeton and Cornell Universities.

‡ The main properties of these quantities have also been rederived in the Bondi–Sachs formalism by van der Burg (1966).

All the aforementioned quantities can be evaluated by performing 2-surface integrals at infinity. A conservation law results when the value of such an integral is independent of the 'time' at which 2-surface is chosen.‡ In the present case, we can express the fact that outgoing waves do, or do not, 'carry away' the quantity in question by performing the integral at the future null infinity of a constant *retarded time* $\mathfrak{u} = \mathfrak{u}_0$ (figure 1) and comparing the result with the integral at a later retarded time $\mathfrak{u} = \mathfrak{u}_1$ ($> \mathfrak{u}_0$). The difference will express the contribution (if any) from the outgoing waves which escape between the two hypersurfaces $\mathfrak{u} =$ const. If this difference necessarily vanishes, we have an absolutely conserved quantity. Otherwise we do not.

A remark concerning the status of incoming waves is appropriate here. The conditions of approach to flatness at infinity that will be adopted are essentially standard ones (Bondi *et al.* 1962; Sachs 1962*a*, *b*; Newman & Penrose 1962; Newman & Unti 1962, cf. also Penrose 1965); which have, on occasion, been interpreted as representing the *absence* of incoming radiation. But incoming *pulse* waves† do not destroy the asymptotic conditions (since the conditions are applied only to future null infinity), nor, moreover, do incoming waves which *die off suitably in time* (cf. also Statchel 1966). It will be important to recognize that such incoming waves are admitted by the formalism, since they will play an important role in the interpretation of the new conserved quantities.

---

† The neutrino field is treated formally as a 'classical' field here (that is, classical apart from the sign ambiguity). It is not clear, therefore, that the results have physical significance for this case.

‡ The familiar form of conservation law arises from a 3-surface integral involving a quantity whose coordinate 4-divergence vanishes. The time-dependence of this integral is then determined by values of components of the quantity (its flux) at infinity, so we would have an absolutely conserved quantity if these components at infinity necessarily vanish. Now, it is generally possible to convert such a 3-surface integral to an explicit 2-surface

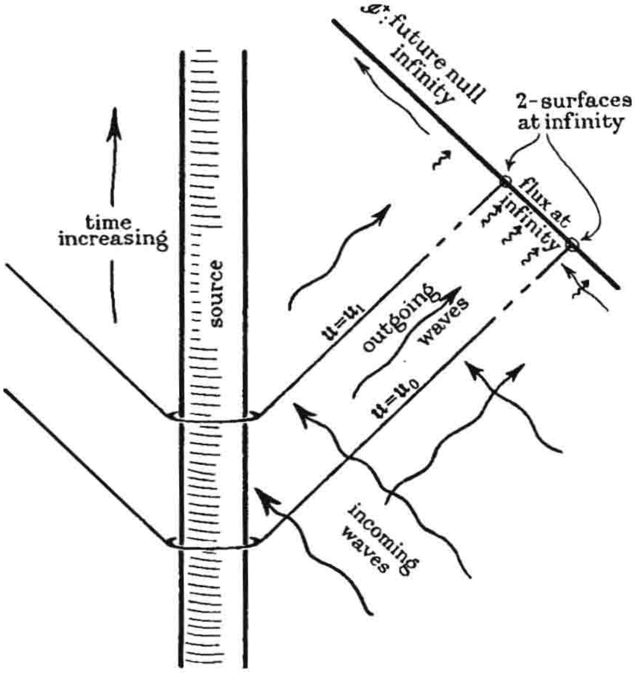

$\mathscr{I}^+$: future null infinity
2-surfaces at infinity
flux at infinity
time increasing
source
$\mathfrak{u}=\mathfrak{u}_1$
outgoing waves
$\mathfrak{u}=\mathfrak{u}_0$
incoming waves

FIGURE 1. Quantities may be defined by 2-surface integrals at infinity on $\mathfrak{u} = \text{const}$. If such a quantity is to contain no contribution from *outgoing* waves, then the integral must be independent of $\mathfrak{u}$.

It will be observed from figure 1, however, that we cannot expect to measure any contribution from incoming waves to such a quantity by examining its $\mathfrak{u}$-dependence. The flux at infinity which shows up in the $\mathfrak{u}$-dependence can involve only outgoing field (see Penrose 1964, 1965). We might envisage measuring the incoming contribution if, instead, we were to use an *advanced* time parameter $\mathfrak{v}$ and define our quantity at past null infinity on $\mathfrak{v} = \text{const}$. This apparently works for energy-momentum, but a curious feature of the new conserved quantities is that

---

integral of a second expression at infinity. Thus, it is really only behaviour at *infinity* that is involved, here, and it is not essential to have started from a quantity with vanishing 4-divergence defined throughout all space-time. In fact, for the new conservation laws described here, no such quantity appears naturally to arise.

† That is, waves for which the disturbance is restricted to lie entirely to the *past* of a null hypersurface which opens out into the past. The existence of such waves would appear to be ensured by consideration of a time-reversed initial value problem. Dispersion effects (i.e. time-reversed 'tails') would then only affect regions to the past of this null hypersurface.

they are *asymmetrical with respect to time.*† Thus, although quantities analogous to the original ones can be defined simply by applying a time-reversal to the definition, these resulting quantities would not, in general, be equal to the original ones. We have, in fact, two sets of 'absolutely' conserved quantites, one referring to the future and the other to the past. (This is assuming that the required boundary conditions hold both in the future and in the past.) It will turn out that, in a sense, we must apparently attribute *some* contribution from incoming waves to the new quantities defined at future null infinity and *some* contribution from outgoing waves to the quantities defined at past null infinity. But this will not show up in the above way and requires somewhat different considerations.

One final remark should be made about conservation laws of this type. This concerns the nature of the retarded time parameter $\mathfrak{u}$. The hypersurfaces $\mathfrak{u} =$ const. are null and they open out into the future. It is customary, and for many purposes extremely convenient, to impose the further condition that the $\mathfrak{u} =$ const. hypersurfaces are, in an appropriate sense 'parallel at infinity' (Bondi *et al.* 1962; Sachs 1962*a*, *b*; Newman & Unti 1962). Then the hypersurfaces can be transformed into each other by an (asymptotic) *translation*. However, not all outgoing null hypersurfaces can be obtained from a given one by such a translation. In general, one requires what is known as a *supertranslation* (Bondi *et al.* 1962; Sachs 1962*a*, *b*; see also Newman & Penrose 1966; Crampin & Foster 1966) which involves, also the asymptotic 'warping' that is generally necessary to transform one null hypersurface into another. From the point of view of the geometry of the space-time there may be no reason, when gravitational radiation is present, to prefer one set of outgoing hypersurfaces to another. Thus, if a quantity can be defined asymptotically in terms of one outgoing null hypersurface, we should expect to be able to refer the definition to *any* other outgoing null hypersurface—not just to those which are time-translations of the first one. We shall call such a quantity absolutely conserved here if its value is *unaltered by supertranslations,* since this expresses the fact that no contribution to the quantity is carried away by the fields between *any* two outgoing null hypersurfaces.

Unfortunately if turns out to be considerably harder to establish the supertranslation invariance of the new quantities than their translation invariance. But the main features of these quantities seem to emerge from the study of their time-translation invariance, so we shall devote most of our attention to a formalism suited to this (§§ 2, 3 and 4). In fact these quantities will also form a representation of the homogeneous Lorentz group (of conformal motions of the 2-sphere at infinity on $\mathfrak{u} =$ const.). Their space-translation invariance will then also follow, since any space-translation can be expressed as a commutator $\mathscr{L}^{-1}\mathscr{T}^{-1}\mathscr{L}\mathscr{T}$ of a time translation $\mathscr{T}$ and a homogeneous Lorentz transformation $\mathscr{L}$. However, we shall also require supertranslation invariance, and a different formalism (Penrose 1965) which gives a somewhat more general viewpoint will be used to treat this (§ 5).

---

† The justification for regarding the energy-momentum expression as describing the 'same' physical quantity whether the expression is used at future or at past null infinity, would lie in the equality of the past limit of the future-measured quantity with the future limit of the past-measured quantity (i.e. both defining the value at I^0; cf. Penrose 1965). Both limits should describe the complete energy-momentum including *all* contributions from gravitational radiation present at any time. These limits are indeed equal if suitable extra boundary conditions are imposed and they are (rather trivially) always equal in the linear theory. But the corresponding 'limits' for the new conserved quantities are only equal to each other in very special circumstances—even in the linear theory.

From a physical point of view, this supertranslation invariance of the new conserved quantities is important because it is required in order to establish a certain (perhaps somewhat surprising) property of asymptotically flat gravitational fields. This is that a stationary field cannot become non-stationary and then stationary again in a finite (retarded) time unless certain parameters, describing the asymmetry of the field (defined in (4·36)), return to their original values. (Whether such a field could return to stationarity even if these parameters do resume their original values is an open question at present.)

To our knowledge, the new conservation laws are not related to any previously known laws. No really adequate physical interpretation of the new quantities is apparent as yet, but at least a tentative description of their meaning can be given. This description rests heavily on their interpretation in the linear theories, where their conservation is, however, essentially trivial, and on their interpretation for static fields. The linear gravitational and Maxwell theories will therefore be analysed in some detail.

In § 2 we summarize some useful information about a differential operator ($\eth$) and a related class of spherical functions which play an important role in the description of the conservation laws. In § 3, we review some earlier work and describe in detail the conservation laws in the Maxwell and linearized Einstein theories giving the interpretation of the conservation laws (now essentially trivial) in terms of incoming waves. In § 4 we derive the conservation under time-translation of the ten quantities for the full non-linear Einstein (vacuum) theory and give their interpretation, in the case of stationary fields, in terms of multipole moments. The alternative approach given in § 5 is based on an intimate connexion between the conserved quantities for each spin and a certain integral expression (Penrose 1963, 1964) which generalizes, to fields of arbitrary spin in flat space-time, Kirchoff's integral for the wave equation. This leads to a proof of the full supertranslation invariance of the new quantities, as required for the discussion of transitions between exactly stationary states (cf. also Penrose 1967). In § 6, there is further discussion of the meaning of the conservation laws.

2. Mathematical preliminaries

In this section we shall review some earlier work (Newman & Penrose 1966; cf. also Goldberg *et al.* 1968) on a useful differential operator which allows the Maxwell and Einstein equations to take a simple form and permits a relatively easy derivation of the conservation laws under time-translation. In addition we shall discuss a certain set of complete functions defined on the sphere for each integer (or half integer), s, called spin s spherical harmonics, which are particularly appropriate for the discussion of the conservation laws.

For simplicity, consider first, ordinary three-dimensional flat space, into which we introduce ordinary spherical polar coordinates r, θ, φ. At each point of each sphere $r =$ const. we introduce the orthonormal triad of vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, where $\mathbf{a}$ and $\mathbf{b}$ are tangent, and $\mathbf{c}$ normal, to the sphere. The vectors $\mathbf{a}$ and $\mathbf{b}$ are defined up to a rotation about $\mathbf{c}$ through an angle χ. Such a rotation is conveniently described by the transformation

$$\mathbf{m}' = \mathrm{e}^{\mathrm{i}\chi}\,\mathbf{m} \tag{2·1}$$

where $\mathbf{m} = 2^{-\frac{1}{2}}(\mathbf{a}+\mathrm{i}\mathbf{b})$. We define a quantity η to be of spin weight s (integer of half-integer) if it transforms under (2·1) by

$$\eta' = \mathrm{e}^{\mathrm{i}s\chi}\,\eta \tag{2·2}$$

(Examples of quantities of spin weights one and minus one, respectively, would be $\mathbf{C}\,.\,\mathbf{m}$ and $\mathbf{C}\,.\,\bar{\mathbf{m}}$, where $\mathbf{C}$ is any ordinary vector.)

Now, for convenience, we specialize the choice of $\mathbf{m}$ so that the vectors $\mathbf{a}$ and $\mathbf{b}$ are tangential, respectively, to the curves $\varphi =$ const. and $\theta =$ const. on each sphere. Then any rotation of the coordinate system (about the origin) must be accompanied by a transformation (2·1). The differential operator $\eth$ defined by

$$\eth\eta = -(\sin\theta)^{s}\left(\frac{\partial}{\partial\theta}+\frac{\mathrm{i}}{\sin\theta}\frac{\partial}{\partial\varphi}\right)\{(\sin\theta)^{-s}\,\eta\} \tag{2·3}$$

(where η has spin weight s) then has the important invariance property of being spin weight unity under the transformations (2·1) induced by rotations of the coordinate system, i.e. (with η as in (2·2)),

$$\eth'\eta' = \mathrm{e}^{\mathrm{i}(s+1)\chi}\,\eth\eta. \tag{2·4}$$

The complex conjugate operator $\bar{\eth}$ defined by

$$\bar{\eth}\eta = -(\sin\theta)^{-s}\left(\frac{\partial}{\partial\theta}-\frac{\mathrm{i}}{\sin\theta}\frac{\partial}{\partial\varphi}\right)\{(\sin\theta)^{s}\,\eta\}$$

similarly has the property

$$\bar{\eth}'\eta' = \mathrm{e}^{\mathrm{i}(s-1)\chi}\,\bar{\eth}\eta,$$

i.e. of being spin weight minus one.

Another important property of $\eth$ and $\bar{\eth}$ is their effect on the spherical harmonics $Y_{l,m}$ $(l = 0, 1, 2, \ldots;\ m = -l, \ldots, +l)$. We can define the spin s spherical harmonics (for integral spin) as follows:

$${}_{s}Y_{l,m} = \left\{\begin{array}{ll} \sqrt{\left\{\dfrac{(l-s)!}{(l+s)!}\right\}}\,\eth^{s}\,Y_{l,m} & (0 \leqslant s \leqslant l), \\ (-1)^{s}\sqrt{\left\{\dfrac{(l+s)!}{(l-s)!}\right\}}\,\bar{\eth}^{-s}\,Y_{l,m} & (-l, \leqslant s \leqslant 0). \end{array}\right\} \tag{2·5}$$

(compare Janis & Newman 1965, Newman & Penrose 1965; Lamb 1965). The ${}_{s}Y_{l,m}$ (which are not defined for $|s| > l$) form a complete orthonormal set for each value of s. We have ${}_{s}\bar{Y}_{l,m} = (-1)^{m+s}\,{}_{-s}Y_{l,-m}$ and

$$\eth\,{}_{s}Y_{l,m} = \sqrt{\{(l-s)\,(l+s+1)\}}\,{}_{s+1}Y_{l,m}, \quad \bar{\eth}\,{}_{s}Y_{l,m} = -\sqrt{\{(l+s)\,(l-s+1)\}}\,{}_{s-1}Y_{l,m}. \tag{2·6}$$

Furthermore, the ${}_{s}Y_{l,m}$ are eigenfunctions, for spin weight s, of the operator $\bar{\eth}\,\eth$, with eigenvalues $-(l-s)\,(l+s+1)$, i.e.

$$\bar{\eth}\eth\,{}_{s}Y_{l,m} = -(l-s)\,(l+s+1)\,{}_{s}Y_{l,m}. \tag{2·7}$$

Note one consequence of (2·6), namely that $\eth$ annihilates ${}_lY_{l,m}$ while $\bar{\eth}$ annihilates ${}_{-l}Y_{l,m}$. More generally, $\eth^p$ [resp. $\bar{\eth}^p$] annihilates any quantity composed only of harmonics ${}_sY_{l,m}$ for which $l-s<p$ [resp. $l+s<p$]. From this we can deduce (integrations being carried out over a sphere $r=$ const.)

$$\int {}_sY_{l,m}\,\eth^{l-s+1}\xi\,\mathrm{d}\omega = 0 \tag{2·8}$$

and

$$\int {}_s\overline{Y}_{l,m}\,\bar{\eth}^{l-s+1}\zeta\,\mathrm{d}\omega = 0 \tag{2·9}$$

where ξ has spin weight $-l-1$ and ζ has spin weight $l+1$ and where

$$\mathrm{d}\omega = \mathrm{d}\theta\sin\theta\,\mathrm{d}\varphi. \tag{2·10}$$

To prove (2·8) and (2·9) we need only integrate repeatedly by parts using

$$\int A\,\eth B\,\mathrm{d}\omega = -\int B\,\eth A\,\mathrm{d}\omega,$$

where the sum of the spin weights of A and B is -1. Alternatively, (2·8) and (2·9) also follow from the orthogonality properties of the ${}_sY_{l,m}$. Another useful relation which follows directly from (2·7), is

$$\int {}_s\overline{Y}_{l,m}\,\bar{\eth}\eth\eta\,\mathrm{d}\omega = -(l-s)(l+s+1)\int {}_s\overline{Y}_{l,m}\,\eta\,\mathrm{d}\omega \tag{2·11}$$

η having spin weight s. (Integrate twice by parts, or else expand η in spin s harmonics.) We note, also, the equation

$$\bar{\eth}\eth\eta - \eth\bar{\eth}\eta = 2s\eta. \tag{2·12}$$

Equations (2·7), (2·8), (2·9), (2·11) will be important in the discussion of the new conservation laws.

3. MAXWELL AND LINEAR GRAVITATIONAL THEORY

In an earlier paper (Newman & Penrose 1962), a tetrad or spinor form of Maxwell's equations was presented; these were further analysed in flat space-time by Janis & Newman (1965), where a special form associated with a special coordinate system was given. These results are now summarized, with but some slight notational changes.†

In flat space-time, a form of null polar coordinates will be introduced, θ and φ being the usual spherical polar angles, as in § 2, but with new coordinates $\mathfrak{r}$ and $\mathfrak{u}$

---

† Agreement with Newman & Penrose (1962) and Exton *et al.* (1968) is being maintained, but to emphasize that changes from Newman & Unti (1962) and Janis & Newman (1965) are involved, the retarded time and radial coordinates, respectively, are being denoted here by $\mathfrak{u}$ and $\mathfrak{r}$, with lower case ψ's and ϕ's for the complex Weyl and Maxwell tensor components, respectively. The translation to the present notation is achieved as follows (with primes indicating the Newman–Unti and Janis–Newman quantities): $\mathfrak{u}' = 2^{\frac{1}{2}}\mathfrak{u}$, $\mathfrak{r}' = 2^{-\frac{1}{2}}\mathfrak{r}$, $l^{\mu\prime} = 2^{\frac{1}{2}}l^{\mu}$, $m^{\mu\prime} = m^{\mu}$, $n^{\mu\prime} = 2^{-\frac{1}{2}}n^{\mu}$; thus, we have $\Psi^{p\prime}_{q} = 2^{-\frac{1}{2}(3+p)}\psi^{p}_{q}$, $\Phi^{p\prime}_{q} = 2^{-\frac{1}{2}(2+p)}\phi^{p}_{q}$, with $\Psi'_{q} = 2^{\frac{1}{2}(2-q)}\psi_{q}$, $\Phi'_{q} = 2^{\frac{1}{2}(1-q)}\phi_{q}$; $\dot{\Psi}^{p\prime}_{q} = 2^{-\frac{1}{2}(4+p)}\dot{\psi}^{p}_{q}$, $\dot{\Phi}^{p\prime}_{q} = 2^{-\frac{1}{2}(3+p)}\dot{\phi}^{p}_{q}$, $\sigma' = 2^{\frac{1}{2}}\sigma$, $\sigma^{0\prime} = 2^{-\frac{1}{2}}\sigma^{0}$, $\dot{\sigma}^{0\prime} = \frac{1}{2}\dot{\sigma}^{0}$, $\ddot{\sigma}^{0\prime} = 2^{-\frac{3}{2}}\ddot{\sigma}^{0}$.

introduced, for which $\mathrm{r}/\sqrt{2}$ measures radial distance and $(2\mathrm{u}+\mathrm{r})/\sqrt{2}$ measures time in the usual system. Thus $\sqrt{2}\,\mathrm{u}$ is the usual retarded time parameter and $\sqrt{2}\,(\mathrm{u}+\mathrm{r})$ is the corresponding advanced time. The hypersurfaces $\mathrm{u} = \text{const.}$ are just the outgoing null cones with vertices on $\mathrm{r} = 0$. The metric is

$$ds^2 = 2\,\mathrm{du}^2+2\,\mathrm{du}\,\mathrm{dr}-\tfrac{1}{2}\mathrm{r}^2(\mathrm{d}\theta^2+\sin^2\theta\,\mathrm{d}\varphi^2) = g_{\mu\nu}\,\mathrm{d}x^\mu\,\mathrm{d}x^\nu. \tag{3·1}$$

with

$$x^0 = \mathrm{u}, \quad x^1 = \mathrm{r}, \quad x^2 = \theta, \quad x^3 = \varphi. \tag{3·2}$$

At any point in space-time a complex tetrad can be chosen in the following manner: l^μ is the outward null vector tangent to the cone $\mathrm{u} = \text{const.}$, n^μ is the inward null vector pointing towards $\mathrm{r} = 0$ and m^μ and $\overline{m}^\mu$ are complex vectors tangent to the two-dimensional sphere defined by constant r and u. ($\overline{m}^\mu$ is the complex conjugate of m^μ.) These vectors in the above null coordinate system are chosen to have the form

$$\left.\begin{aligned} &l^\mu = \delta^\mu_1, \quad n^\mu = \delta^\mu_0-\delta^\mu_1,\\ &m^\mu = \frac{1}{\mathrm{r}}\left(\delta^\mu_2+\frac{\mathrm{i}}{\sin\theta}\,\delta^\mu_3\right), \quad \overline{m}^\mu = \frac{1}{\mathrm{r}}\left(\delta^\mu_2-\frac{\mathrm{i}}{\sin\theta}\,\delta^\mu_3\right).\end{aligned}\right\} \tag{3·3}$$

In order to describe the electromagnetic field we use the three complex tetrad components of the Maxwell field tensor $F_{\mu\nu}$, defined by

$$\left.\begin{aligned} \phi_0 &= F_{\mu\nu}l^\mu m^\nu,\\ \phi_1 &= \tfrac{1}{2}F_{\mu\nu}(l^\mu n^\nu+\overline{m}^\mu m^\nu),\\ \phi_2 &= F_{\mu\nu}\overline{m}^\mu n^\nu.\end{aligned}\right\} \tag{3·4}$$

In terms of the triad of vectors $\mathbf{m}$, $\bar{\mathbf{m}}$ and $\mathbf{c}$ of § 2 and (3·3), we have, in fact

$$\left.\begin{aligned} \phi_0 &= 2^{-\frac{1}{2}}(\mathbf{E}+\mathrm{i}\mathbf{B})\cdot\mathbf{m},\\ \phi_1 &= \tfrac{1}{2}(\mathbf{E}+\mathrm{i}\mathbf{B})\cdot\mathbf{c},\\ \phi_2 &= -\,2^{-\frac{1}{2}}(\mathbf{E}+\mathrm{i}\mathbf{B})\cdot\bar{\mathbf{m}}.\end{aligned}\right\} \tag{3·5}$$

From this, or directly from (3·4), we see that ϕ_0, ϕ_1 and ϕ_2 are respectively of spin weights $1, 0$ and -1. (The appropriate type of spherical harmonic for the description of ϕ_0, ϕ_1 and ϕ_2, respectively, is thus given by ${}_1Y_{l,m}$ ${}_0Y_{l,m}$ and ${}_{-1}Y_{l,m}$.)

The Maxwell equations, written in terms of the ϕ's, the null polar coordinates and the differential operator $\eth$, have the form

$$\left(\frac{\partial}{\partial\mathrm{r}}+\frac{2}{\mathrm{r}}\right)\phi_1+\frac{1}{\mathrm{r}}\bar{\eth}\phi_0 = 0, \tag{3·6}$$

$$\left(\frac{\partial}{\partial\mathrm{r}}+\frac{1}{\mathrm{r}}\right)\phi_2+\frac{1}{\mathrm{r}}\bar{\eth}\phi_1 = 0, \tag{3·7}$$

$$\left(\frac{\partial}{\partial\mathrm{u}}-\frac{\partial}{\partial\mathrm{r}}-\frac{1}{\mathrm{r}}\right)\phi_0+\frac{1}{\mathrm{r}}\,\eth\phi_1 = 0, \tag{3·8}$$

$$\left(\frac{\partial}{\partial\mathrm{u}}-\frac{\partial}{\partial\mathrm{r}}-\frac{2}{\mathrm{r}}\right)\phi_1+\frac{1}{\mathrm{r}}\,\eth\phi_2 = 0. \tag{3·9}$$

We shall be concerned with solutions which satisfy an appropriate asymptotic boundary condition. This condition will admit *all* sufficiently smooth *retarded* fields and, in addition, fields with *incoming* waves which *die off* suitably. Using the integer N to measure the degree of smoothness required, we demand:

$$\phi_0 = \sum_{n=0}^{N} \frac{\phi_0^n(\mathrm{u}, \theta, \varphi)}{\mathrm{r}^{3+n}} + \mathrm{o}(\mathrm{r}^{-3-N}), \tag{3·10}$$

where we must also assume that a sufficient number of r, θ, φ and u derivatives of the expression (3·10) also hold (cf. Newman & Penrose 1962; Janis & Newman 1965). Then (3·10) also governs the behaviour of ϕ_1 and ϕ_2. (In the conformal approach of Penrose 1965, the condition (3·10) arises if we assume C^N differentiability of the transformed electromagnetic field at $\mathscr{I}^+$.)

Equations (3·6), (3·7) can be integrated, yielding

$$\phi_1 = \frac{\phi_1^0}{\mathrm{r}^2} + \sum_{n=0}^{N} \frac{\bar{\eth}\phi_0^n}{(n+1)\,\mathrm{r}^{3+n}} + \mathrm{o}(\mathrm{r}^{-3-N}), \tag{3·11}$$

$$\phi_2 = \frac{\phi_2^0}{\mathrm{r}} + \frac{1}{\mathrm{r}^2}\,\bar{\eth}\phi_1^0 + \sum_{n=0}^{N} \frac{\bar{\eth}^2\phi_0^n}{(n+1)\,(n+2)\,\mathrm{r}^{3+n}} + \mathrm{o}(\mathrm{r}^{-3-N}). \tag{3·12}$$

When these are substituted into (3·8), (3·9), we obtain, with the *dot* denoting $\partial/\partial \mathrm{u}$:

$$\dot{\phi}_1^0 = -\eth\phi_2^0, \quad \dot{\phi}_0^0 = -\eth\phi_1^0 \tag{3·13}$$

and (cf. (2·12))

$$(n+1)\,\dot{\phi}_0^{n+1} = -\{\bar{\eth}\eth\phi_0^n + (n+3)\,n\phi_0^n\} \quad (N > n \geqslant 0). \tag{3·14}$$

Before deriving the new conservation laws, we first exhibit the law of conservation of charge in this formalism. By multiplying the first equation of (3·13) by ${}_0Y_{00}$ (this is done for the sake of analogy with later work; in fact ${}_0Y_{0,0} = (4\pi)^{-\frac{1}{2}}$) and integrating over the sphere, we obtain

$$\frac{\mathrm{d}}{\mathrm{du}}\int {}_0Y_{0,0}\,\phi_1^0\,\mathrm{d}\omega = -\int {}_0Y_{0,0}\,\eth\phi_2^0\,\mathrm{d}\omega = 0 \tag{3·15}$$

by (2·8). Thus, the quantity $\int {}_0Y_{0,0}\,\phi_1^0\,\mathrm{d}\omega$ is constant (independent of u). This quantity is proportional to the total electric charge (plus i times 'magnetic charge') (cf. Janis & Newman 1965).

Similarly, to obtain the new conserved quantities, we can multiply (3·14) by ${}_1\overline{Y}_{n-k+1,m}$ and integrate over the sphere (cf. (2·11)):

$$\begin{aligned}(n+1)\frac{\mathrm{d}}{\mathrm{du}}\int {}_1Y_{n-k+1,m}\,\phi_0^{n+1}\,\mathrm{d}\omega &= -\int {}_1\overline{Y}_{n-k+1,m}(\bar{\eth}\eth\phi_0^n + (n+3)\,n\phi_0^n)\,\mathrm{d}\omega \\ &= -k(2n-k+3)\int {}_1\overline{Y}_{n-k+1,m}\,\phi_0^n\,\mathrm{d}\omega.\end{aligned} \tag{3·16}$$

Thus if

$$F_m^{n,k} = \int {}_1\overline{Y}_{n-k+1,m}\,\phi_0^{n+1}\,\mathrm{d}\omega, \tag{3·17}$$

where m, n and k are integers with $n \geqslant k \geqslant 0$, $|m| \leqslant n-k+1$, we have

$$\dot{F}_m^{n,k} = -\frac{k(2n-k+3)}{n+1}\,F_m^{n-1,k-1}; \quad \dot{F}_m^{n,0} = 0. \tag{3·18}$$

This gives us a hierarchy of *conserved quantities* $F^{n,0}_m$ (provided $N \geqslant n+1$). The only ones of these which will remain conserved in the non-linear Einstein–Maxwell theory, however, are those given by $n = 0$, $m = -1, 0, 1$. We shall thus concern ourselves mainly with these three. We shall write $F_m = F^{0,0}_m$.

From the fact (cf. Janis & Newman 1965; Lamb 1966) that

$$\phi^n_0 = \sum_{l=n+1}^{\infty} \alpha^n_{l,m}\,{}_1Y_{l,m},$$

for all (sufficiently smooth) outgoing multipole solutions, it follows that all the $F^{n,k}_m$ are zero in the absence of incoming waves. It turns out that $F^{n,k}_m$ measure asymptotic properties (for large values of the advanced time coordinate $\mathfrak{v} = \mathfrak{u}+\mathfrak{r}$) of the time profile of the *incoming* waves. We shall analyse, in detail, the case of an incoming dipole field with the dipole axis along the z-direction.

The field of the advanced dipole described with retarded coordinates is

$$\left.\begin{aligned} \phi_0 &= -\sin\theta\,\frac{\partial^2}{\partial\mathfrak{r}^2}\left(\frac{\alpha}{\mathfrak{r}}\right), \\ \phi_1 &= -2\cos\theta\,\frac{\partial}{\partial\mathfrak{r}}\left(\frac{\alpha}{\mathfrak{r}^2}\right), \\ \phi_2 &= +2\sin\theta\,\frac{\alpha}{\mathfrak{r}^3}, \end{aligned}\right\} \tag{3·19}$$

where $\alpha = \alpha(\mathfrak{u}+\mathfrak{r})$ is an arbitrary complex function of the advanced time $\mathfrak{u}+\mathfrak{r}$ and represents the electric dipole moment plus i times the magnetic dipole moment. This solution can be easily verified by direct substitution in the Maxwell equations (3·6)–(3·9). To satisfy (3·10) we must assume (taking $N = 1$):

$$\alpha = \beta(\mathfrak{u}+\mathfrak{r})^{-1} + o((\mathfrak{u}+\mathfrak{r})^{-1})$$

(β = const.), and that at least two $\mathfrak{r}$-derivatives commute with the order symbol. Substituting into (3·19) and using $(\mathfrak{u}+\mathfrak{r})^{-1} = \mathfrak{r}^{-1} - \mathfrak{u}\mathfrak{r}^{-2} + \mathfrak{u}^2\mathfrak{r}^{-3} - \ldots$, we get

$$\phi_0 = -6\beta\mathfrak{r}^{-4}\sin\theta + o(\mathfrak{r}^{-4}), \tag{3·20}$$

so, by (3·10), $\phi^1_0 = -6\beta\sin\theta$ giving (cf. (3·17))

$$F_0 = -8(3\pi)^{\frac{1}{2}}\beta, \quad F_{-1} = F_1 = 0.$$

Thus, the F_m are defined simply by the coefficient of the $(\mathfrak{u}+\mathfrak{r})^{-1}$ term in the asymptotic expansion of the dipole incoming field and are hence of apparently little fundamental importance in flat space-time. A similar statement can be made about the other $F^{n,k}_m$.

The results presented above for the Maxwell theory are easily extended to the case of linearized gravitational theory. The five complex functions

$$\left.\begin{aligned} \psi_0 &= -C_{\mu\nu\rho\sigma}l^\mu m^\nu l^\rho m^\sigma, \\ \psi_1 &= -C_{\mu\nu\rho\sigma}l^\mu n^\nu l^\rho m^\sigma, \\ \psi_2 &= -C_{\mu\nu\rho\sigma}\overline{m}^\mu n^\nu l^\rho m^\sigma, \\ \psi_3 &= -C_{\mu\nu\rho\sigma}\overline{m}^\mu n^\nu l^\rho n^\sigma, \\ \psi_4 &= -C_{\mu\nu\rho\sigma}\overline{m}^\mu n^\nu \overline{m}^\rho n^\sigma, \end{aligned}\right\} \tag{3·21}$$

where $C_{\mu\nu\rho\sigma}$ is the linearized Weyl tensor (Sachs & Bergmann 1958), are the analogues of the ϕ's in Maxwell theory. The differential equations that they satisfy are the linearized Bianchi identities (cf. also Newman & Penrose 1962; Janis & Newman 1965; Penrose 1965). Just as the $F_{\mu\nu}$ corresponds to $(\mathbf{E}, \mathbf{B})$, the Weyl tensor corresponds to two trace-free symmetric, 3-dimensional tensors C_{km} and E_{km}, where C_{km} is $C_{\alpha\kappa\beta\mu}(l^\alpha+n^\alpha)(l^\beta+n^\beta)$ and E_{km} is $C^*_{\alpha\kappa\beta\mu}(l^\alpha+n^\alpha)(l^\beta+n^\beta)$ (the asterisk denoting dualization). Putting

$$Z_{kj} = \tfrac{1}{2}C_{kj} + \frac{\mathrm{i}}{2}E_{kj},$$

we get

$$\psi_0 = -Z_{kj}m_k m_j, \quad \psi_1 = 2^{\frac{1}{2}}Z_{kj}c_k m_j, \quad \psi_2 = Z_{kj}m_k\overline{m}_j,$$
$$\psi_3 = -2^{\frac{1}{2}}Z_{kj}c_k\overline{m}_j, \quad \psi_4 = -Z_{kj}\overline{m}_k\overline{m}_j,$$

where c_j, m_j and $\overline{m}_j$ are the complex triad of §2. Note that ψ_n has spin weight $2-n$ $(n = 0, 1, \ldots, 4)$.

As with the Maxwell theory, we shall be concerned with solutions satisfying a boundary condition at infinity which admits all sufficiently smooth retarded fields and also appropriately tailing off incoming fields. In fact we require behaviour of the form

$$\psi_0 = \sum_{n=0}^{N} \frac{\psi_0^n}{\mathfrak{r}^{5+n}} + o(\mathfrak{r}^{-5-N}), \tag{3·22}$$

where sufficiently many $\mathfrak{r}$, θ, φ and $\mathfrak{u}$ derivatives of the expression hold. (In the conformal approach, (3·22) follows from C^N differentiability of the transformed field at $\mathscr{I}^+$.) The 'Bianchi identities' can then be integrated to yield (Newman & Penrose 1962; Janis & Newman 1965):

$$\left.\begin{aligned}
\psi_1 &= \frac{\psi_1^0}{\mathfrak{r}^4} + o(\mathfrak{r}^{-4}),\\
\psi_2 &= \frac{\psi_2^0}{\mathfrak{r}^3} + o(\mathfrak{r}^{-3}),\\
\psi_3 &= \frac{\psi_3^0}{\mathfrak{r}^2} + o(\mathfrak{r}^{-2}),\\
\psi_4 &= \frac{\psi_4^0}{\mathfrak{r}} + o(\mathfrak{r}^{-1}).
\end{aligned}\right\} \tag{3·23}$$

The time-development equations (the second set of the Bianchi identities) take the following form when (3·22), (3·23) and $\eth$ are used:

$$\dot{\psi}_3^0 = -\eth\psi_4^0, \tag{3·24}$$

$$\dot{\psi}_2^0 = -\eth\psi_3^0, \tag{3·25}$$

$$\dot{\psi}_1^0 = -\eth\psi_2^0, \tag{3·26}$$

$$\dot{\psi}_0^0 = -\eth\psi_1^0, \tag{3·27}$$

$$(n+1)\,\dot{\psi}_0^{n+1} = -\{\bar{\eth}\eth + (n+5)n\}\,\psi_0^n \quad (N > n \geqslant 0). \tag{3·28}$$

The conservation of mass follows from (3·25) in exactly the same way that conservation of charge followed from (3·13)

$$\frac{\mathrm{d}}{\mathrm{d}u}\int {}_0Y_{0.0}\,\psi_2^0\,\mathrm{d}\omega = -\int {}_0Y_{0.0}\,\eth\psi_3^0\,\mathrm{d}\omega = 0, \tag{3·29}$$

by (2·8). The total mass is proportional to $\int {}_0Y_{0,0}\,\psi_2^0\,\mathrm{d}\omega$. The conservation laws of linear and angular momentum are slightly harder to prove, as the proof depends on the existence of the linearized metric tensor as well as the Weyl tensor. We shall not give the argument here (but see Sachs & Bergmann 1958; Janis & Newman 1965). The essential significance, in this context, of the existence of the linearized metric is that it implies that the dipole parts of ψ_3^0 and ψ_2^0 are, respectively, purely 'magnetic' and purely 'electric' (see Newman & Penrose 1966). The linear momentum vector is given by

$$\int {}_0Y_{1,m}\,\mathrm{Re}\,(\psi_2^0)\,\mathrm{d}\omega\;\left(=\int {}_0Y_{1,m}\psi_2^0\,\mathrm{d}\omega\right)$$

and the angular momentum vector by

$$\int {}_0Y_{1,m}\,\mathrm{Im}\,(\bar{\eth}\psi_1^0)\,\mathrm{d}\omega \quad (m=-1,0,1).$$

The mass and momentum conservation laws for the full Einstein theory are given in the next section.

The *new conservation laws* in the linear theory can be obtained from (3·28) by multiplying both sides by ${}_2\overline{Y}_{n-k+2,m}$ and integrating over the sphere:

$$\begin{aligned}(n+1)\frac{\partial}{\partial \mathrm{u}}\int {}_2\overline{Y}_{n-k+2,m}\,\psi_0^{n+1}\,\mathrm{d}\omega &= -\int {}_2\overline{Y}_{n-k+2,m}\{\bar{\eth}\eth\psi_0^n+(n+5)\,n\psi_0^n\}\,\mathrm{d}\omega\\ &= -\int {}_2\overline{Y}_{n-k+2,m}\,k(2n-k+5)\psi_0^n\,\mathrm{d}\omega,\end{aligned}$$

see (2·11), whence

$$\dot{G}_m^{n,k} = -\frac{k(2n-k+5)}{n+1}\,G_m^{n-1,k-1};\quad \dot{G}_m^{n,0}=0 \tag{3·30}$$

(n, k and m being integers with $n\geqslant k\geqslant 0$, $|m|\leqslant n-k+2$), with

$$G_m^{n,k} = \int {}_2\overline{Y}_{n-k+2,m}\,\psi_0^{n+1}\,\mathrm{d}\omega. \tag{3·31}$$

We shall concern ourselves mainly with the five quantities $G_m = G_m^{0,0}$.

By performing a time inversion on a retarded quadrupole solution (cf. Janis & Newman 1965) we obtain an advanced quadrupole field, the ψ_0 of which can be put in the form

$$\psi_0 = \frac{\partial^4}{\partial \mathrm{r}^4}\left(\frac{P(\mathrm{u}+\mathrm{r})}{\mathrm{r}}\right)\sin^2\theta, \tag{3·32}$$

where the real and imaginary parts of P are proportional to the incoming mass quadrupole and angular momentum quadrupole moments. In order to compare (3·32) with (3·22), we assume that for large values of $\mathrm{u}+\mathrm{r}$ we have

$$P = (\mathrm{u}+\mathrm{r})^{-1}R+(\mathrm{u}+\mathrm{r})^{-2}S+\mathrm{o}((\mathrm{u}+\mathrm{r})^{-2}) \tag{3·33}$$

(R and S being constants) and that at least four derivatives commute with the order symbol. Then for large r, $P = R\mathrm{r}^{-1}+(S-R\mathrm{u})\mathrm{r}^{-2}+\mathrm{o}(\mathrm{r}^{-2})$ which, when substituted into (3·32) yields

$$\psi_0 = 120R\sin^2\theta\,.\,\mathrm{r}^{-6}+360(S-R\mathrm{u})\sin^2\theta\,.\,\mathrm{r}^{-7}+\mathrm{o}(\mathrm{r}^{-7}), \tag{3·34}$$

i.e. the conserved quantity G_0 is proportional to R, and again appears to be devoid of physical interest in the linear theory, just indicating the existence of an incoming quadrupole field with some particular asymptotic shape. (Note that the R appears as the coefficient of u in the r^{-7} term of the expansion. This will play a role in the interpretation of the conservation law in the non-linear theory.)

4. NON-LINEAR GRAVITATION THEORY

In the preceding section, it was shown that there exists, in the linear theory of gravitation, an infinite number of conserved quantities $G_m^{n,0}$ (cf. (3·31)), all associated with the presence of incoming radiation. (If past light-cones had been used, the conserved quantities would be associated with outgoing radiation.) In the full non-linear theory of gravity the method of proof of the conservation laws will break down, except (as we shall see shortly) in the case of the five G_m given by (4·19). It thus appears likely that in the full theory, there are only these five complex absolutely conserved quantities, although it is conceivable that others might be constructible with the use of an inverse operation $\eth^{-1}$ (compare Exton *et al.* 1968). In this section we shall obtain the time-translation invariance (u-independence) of the G_m in the full theory and show why their interpretation must now be radically altered, i.e. they need not be associated with incoming fields.

We choose coordinates u, r, θ, φ so that $\sqrt{2}\,u$ is a standard Bondi-type 'retarded time' parameter, u = const. being outgoing null hypersurfaces 'parallel at infinity'; θ, φ are standard spherical polar coordinates for the sphere at infinity on each hypersurface; $r\sqrt{2}$ is a standard affine parameter for each generator u, θ, φ = const. (cf. Janis & Newman 1965). We use a standard null tetrad system $l^\mu, n^\mu, m^\mu, \overline{m}^\mu$ parallelly propagated along these generators, with l^μ tangent to the generators, and m^μ related to the θ, φ coordinates as in § 2. We shall assume:

$$\psi_0 = \psi_0^0 r^{-5} + \psi_0^1 r^{-6} + \psi_0^2 r^{-7} + o(r^{-7}) \tag{4·1}$$

and also that a sufficient number of u, θ, φ, r derivatives of this expression also hold. These assumptions are essentially equivalent to those of Bondi *et al.* (1962) and Sachs (1962*a*) except that we do not require complete analytic behaviour in r (which would amount to a full power series assumed in place of (4·1)). In the conformal approach to asymptotic flatness (Penrose 1965) our assumptions would follow from the C^5 differentiability of the conformal factor and transformed metric at $\mathscr{I}^+$. The present assumptions are slightly stronger than have been used in some earlier papers (Newman & Penrose 1962: $\psi_0 = O(r^{-5})$; Newman & Unti 1962: $\psi_0 = \psi_0^0 r^{-5} + O(r^{-6})$) but the higher order terms will be needed here. From the results of these papers:

$$\left.\begin{aligned}\psi_1 &= \psi_1^0 r^{-4} + O(r^{-5}),\\ \psi_2 &= \psi_2^0 r^{-3} + O(r^{-4}),\\ \psi_3 &= \psi_3^0 r^{-2} + O(r^{-3}),\\ \psi_4 &= \psi_4^0 r^{-1} + O(r^{-2}).\end{aligned}\right\} \tag{4·2}$$

For the time-development of the quantities in (4·1) and (4·2) we need to consider the leading term $\sigma^0 = \sigma^0(u, \theta, \varphi)$ in the *shear* σ of the null congruence l^μ:

$$\sigma = l_{\mu;\nu} m^\mu m^\nu = \sigma^0 r^{-2} + O(r^{-4}). \tag{4·3}$$

Unlike the case of the linear theory, it is now not possible to choose $\sigma^0 = 0$ everywhere if there is outgoing radiation. In fact (see Bondi *et al.* 1962; Sachs 1962*a*; Newman & Unti 1962)

$$\ddot{\sigma}^0 = -\bar{\psi}_4^0, \tag{4·4}$$

so that σ^0 is, effectively, a second time-integral of the radiation field. Furthermore,

$$\psi_3^0 = \eth\dot{\bar{\sigma}}^0. \tag{4·5}$$

We shall also require the equation

$$\psi_2^0 - \bar{\psi}_2^0 = \bar{\sigma}^0\dot{\sigma}^0 - \sigma^0\dot{\bar{\sigma}}^0 + \bar{\eth}^2\sigma^0 - \eth^2\bar{\sigma}^0. \tag{4·6}$$

The Bianchi identities give

$$\dot{\psi}_3^0 = -\eth\psi_4^0, \tag{4·7}$$

$$\dot{\psi}_2^0 = -\eth\psi_3^0 + \sigma^0\psi_4^0, \tag{4·8}$$

$$\dot{\psi}_1^0 = -\eth\psi_2^0 + 2\sigma^0\psi_3^0, \tag{4·9}$$

$$\dot{\psi}_0^0 = -\eth\psi_1^0 + 3\sigma^0\psi_2^0. \tag{4·10}$$

all these equations are obtained from Newman & Unti (1962) by translation into the $\eth$ notation. Note that from (4·8), (4·4) and (4·5) we get

$$\dot{\psi}_2^0 = -\eth^2\dot{\bar{\sigma}}^0 - \sigma^0\ddot{\bar{\sigma}}^0. \tag{4·11}$$

Further calculations yield†

$$\dot{\psi}_0^1 = -\bar{\eth}\eth\psi_0^0 - 4\bar{\eth}(\sigma^0\psi_1^0), \tag{4·12}$$

$$\dot{\psi}_0^2 = -\tfrac{1}{2}(\bar{\eth}\eth\psi_0^1 + 6\psi_0^1) + 5\{(\psi_1^0)^2 - \psi_0^0\psi_2^0 - \tfrac{1}{2}\psi_0^0\bar{\psi}_2^0\} + E, \tag{4·13}$$

where

$$\begin{aligned} E = {} & 5\bar{\eth}\sigma^0\bar{\eth}\psi_0^0 + 3\eth\bar{\sigma}^0\eth\psi_0^0 + \tfrac{5}{2}\sigma^0\bar{\eth}^2\psi_0^0 + \tfrac{5}{2}\psi_0^0\eth^2\bar{\sigma}^0 + \tfrac{1}{2}\bar{\sigma}^0\eth^2\psi_0^0 - \tfrac{9}{2}\sigma^0\bar{\sigma}^0\eth\psi_1^0 \\ & - 12\sigma^0\psi_1^0\eth\bar{\sigma}^0 - 2\bar{\sigma}^0\psi_1^0\eth\sigma^0 + \tfrac{15}{2}\bar{\sigma}^0(\sigma^0)^2\psi_2^0 + \tfrac{5}{2}\sigma^0\dot{\bar{\sigma}}^0\psi_0^0. \end{aligned} \tag{4·14}$$

(The important thing about E, here, is that it vanishes when $\sigma^0 = 0$.)

Before considering the new conservation laws in detail, we note, for comparison, the Bondi–Sachs mass-momentum law‡ for which we put

$$P_{n,m} = -\frac{1}{\sqrt{(32\pi)}}\int {}_0Y_{n,m}(\sigma^0\dot{\bar{\sigma}}^0 + \psi_2^0)\,\mathrm{d}\omega, \tag{4·15}$$

$1 \geqslant n \geqslant |m| \geqslant 0$. Then,

$$(P_{0,0},\ 6^{-\frac{1}{2}}(P_{1,1} - P_{1,-1}),\ \ \mathrm{i}6^{-\frac{1}{2}}(P_{1,1} + P_{1,-1}),\ \ 3^{-\frac{1}{2}}P_{1,0}) \tag{4·16}$$

is the Bondi–Sachs total energy-momentum vector (in an 'asymptotic' orthonormal frame) at the retarded time u, with $P_{0,0}$ being the total mass. In fact, this vector is real, even though $\sigma^0\dot{\bar{\sigma}}^0 + \psi_2^0$ is complex, as follows from (4·6), (2·8) and (2·9). The u-dependence is given by

---

† We are deeply indebted to Dr T. Unti for his aid in performing the very lengthy calculations needed here.

‡ Although Bondi *et al.* (1962) and Sachs (1962*a*) were originally explicitly concerned with mass only, the behaviour under homogeneous Lorentz transformations implies that the momentum law is also involved (cf. Sachs 1962*b*). The momentum carried by gravitational radiation is more explicitly considered by Bonnor & Rotenberg (1966); see also Penrose (1964, 1967), Winicour (1967).

$$\dot{P}_{n,m} = -\frac{1}{\sqrt{(32\pi)}}\int {}_0Y_{n,m}(|\dot{\sigma}^0|^2 - \eth^2\dot{\bar{\sigma}}^0)\,\mathrm{d}\omega = -Q_{n,m} \tag{4·17}$$

where

$$Q_{n,m} = \frac{1}{\sqrt{(32\pi)}}\int {}_0Y_{n,m}|\dot{\sigma}^0|^2\,\mathrm{d}\omega \tag{4·18}$$

(see (4·11), (2·8)). The 4-vector corresponding to $Q_{n,m}$ as in (4·16) is *time-like* and *future-pointing* (or zero) as follows from the fact that $|\dot{\sigma}^0|^2$ is non-negative and the fact that the 4-vector corresponding to ${}_0Y_{n,m}$ is null and future-pointing for each θ, φ. The total energy-momentum carried away by the outgoing gravitational radiation between u and u + du is represented by $Q_{n,m}\,\mathrm{du}$. Note that, in particular, the 'Lorentz invariant' Bondi–Sachs *rest-mass* $M = (P_{0,0}^2 - \frac{1}{3}P_{1,0}^2 + \frac{2}{3}P_{1,1}P_{1,-1})^{\frac{1}{2}}$ cannot increase with u and decreases if and only if $\dot{\sigma}^0 \neq 0$.

In fact (4·17) can be generalized to give an expression for the difference between the $P_{n,m}$ evaluated on *any* two outgoing null hypersurfaces $\mathcal{N}, \mathcal{N}'$—which need *not* both be given by u = const. in one standard coordinate system (i.e. they may be obtained, one from the other, by a supertranslation.) The result (obtained in Penrose 1964, 1967; Winicour 1967) is that this difference can again be expressed as an integral of $|\dot{\sigma}^0|^2$. (We do not give the proof, but merely indicate the result here since it will be required shortly in what follows.) Thus, if $\dot{\sigma}^0 = 0$, the $P_{n,m}$ must be the same whether evaluated on $\mathcal{N}$ or on $\mathcal{N}'$. (This assumes that the same, i.e. 'supertranslated', angular coordinates θ, φ are being used on $\mathcal{N}$ and $\mathcal{N}'$. Otherwise the way in which the $P_{n,m}$ components transform among themselves under Lorentz transformation must be taken into account.) Furthermore, if $\mathcal{N}'$ lies (asymptotically) entirely to the future of $\mathcal{N}$, then the rest-mass M, evaluated at $\mathcal{N}$ must exceed that evaluated at $\mathcal{N}'$ *unless* $\dot{\sigma}^0$ vanishes everywhere between $\mathcal{N}$ and $\mathcal{N}'$.

We now consider the '*absolute*' conservation law $\dot{G}_m = 0$, where

$$G_m = \int {}_2\bar{Y}_{2,m}\,\psi_0^1\,\mathrm{d}\omega \quad (m = -2, -1, \ldots, 2). \tag{4·19}$$

This time-independence follows at once from (2·9) since by (4·12), ψ_0^1 has the form $\bar{\eth}A$, with A of spin weight 3. The method used here does not generalize to the ψ_0^2 equation and it would appear that there are no further conserved quantities of this type (that is, which do not involve $\eth^{-1}$).

The first impression of the authors had been that, as in the linear theory, a non-vanishing G_m indicated the existence of an incoming quadrupole radiation field; however, it is here that the non-linear theory presents us with an entirely new situation. That a non-vanishing of G_m does *not* represent incoming radiation alone in the full theory, can be brought out most clearly by an examination of the stationary vacuum solutions. In such cases incoming radiation is absent (according to any reasonable definition) but the G_m turn out to be non-vanishing, in general, and, moreover, expressible as certain interesting combinations of the mass, dipole and quadrupole moments. Since a large part of the discussion of the meaning of the G_m in the full theory hinges on this feature of stationary fields, we shall devote the rest of this section almost entirely to a discussion of such fields. We shall have to pursue the discussion with some care, as there are subtleties involved in connexion with the choice of coordinate system.

Suppose, then, that the field is stationary; that is, there exists a time-like Killing vector field. We shall first show that it is possible to choose our standard coordinate system to that the Killing direction is given by $\partial/\partial u$—that is to say, the dot derivative of every geometrical quantity vanishes—and, furthermore, so that $\sigma^0 = 0$. Consider the outgoing null hypersurface $\mathscr{N}$, defined by $u = 0$ of some given standard coordinate system. The points of $\mathscr{N}$ are labelled by standard coordinates r, θ, φ. Now, by the action of the Killing field, $\mathscr{N}$ is carried into a continuous sequence of other outgoing null hypersurfaces (not necessarily given by $u = \text{const.}$). Let $\mathscr{N}'$ be one of these and suppose that the point $(0, r, \theta, \varphi)$ of $\mathscr{N}$ is carried into the point $(u_0', r', \theta', \varphi')$ of $\mathscr{N}'$. This can be achieved by a unique Bondi–Metzner–Sachs transformation (cf. Bondi *et al.* 1962; Sachs 1962*a*, *b*) which might involve a Lorentz velocity transformation, a spatial rotation and a supertranslation. We can, in fact, eliminate the possibility of a velocity transformation by choosing the 'time axis' of our coordinate system to be parallel to the Bondi–Sachs energy-momentum vector (4·16) at $u = 0$; that is to say, $P_{1,1} = P_{1,0} = P_{1,-1} = 0$ at $u = 0$. For, it is a consequence of the stationarity that the rest-mass M is the same evaluated on $\mathscr{N}'$ as on $\mathscr{N}$. Thus $\dot{\sigma}^0 = 0$ between $\mathscr{N}$ and $\mathscr{N}'$, so the $P_{n,m}$ must be the same evaluated on $\mathscr{N}'$ as on $\mathscr{N}$ (see discussion above): the Bondi–Sachs energy-momentum vector is constant and carried into itself by the action of the Killing field. Furthermore, no spatial rotation can be necessarily involved either, since otherwise we would have a periodic asymptotically flat gravitational field† in violation of a theorem of Papapetrou (1957); cf. also Sachs 1962*a*). (A Killing field involving a rotation at infinity would become spacelike at large distances in any case.) Thus we have $\theta = \theta'$, $\varphi = \varphi'$ and the transformation from $\mathscr{N}$ to $\mathscr{N}'$ is, at worst, a supertranslation: $u' = u + \alpha(\theta, \varphi)$ (with $\alpha = u_0'$). In fact, it must be a (time-) translation since we have $\sigma^0 = \sigma^{0\prime}$ (where $\sigma^{0\prime}$ defines the asymptotic shear of $\mathscr{N}'$ at (θ, φ) in the same way that σ^0 refers to $\mathscr{N}$), whence $\eth^2\alpha = 0$ (cf. Sachs 1962*a*; Newman & Penrose 1966), which is the condition for the supertranslation to specialize to a translation. For $\mathscr{N}'$ lying infinitesimally to the future of $\mathscr{N}$, this shows that the Killing field must be asymptotically a time-translation. Thus, if we now choose r, θ, φ, to be standard coordinates on $\mathscr{N}$ compatible with the 'time-axis' being the asymptotic direction of the Killing field, and define coordinates away from $\mathscr{N}$ so that $\partial/\partial u$ gives us the Killing vector (with $\mathscr{N}$ still being $u = 0$), then we shall have set up a standard coordinate system as required. (In fact, this choice of coordinates still agrees with the above condition $P_{1,1} = P_{1,0} = P_{1,-1} = 0$, as follows from (4·15), with $\dot{\sigma}^0 = 0$, by (4·21) below.)

With these coordinates, all the dotted quantities will vanish. In particular, since $\dot{\psi}_1^0 = 0$, we have, from (4·9), (4·5),

$$\eth\psi_2^0 = 0 \tag{4·20}$$

from which it follows (cf. Newman & Penrose 1966) that ψ_2^0 is constant in θ and φ

---

† Of course, the presence of a Killing field involving a rotation at infinity (such as is present for the Schwarzschild, Kerr and Weyl–Levi-Civita solutions) is not being ruled out here. It is just that such a Killing field cannot be the *only* one present for a stationary solution.

as well as in u. Hence, from (4·15), $P_{0,0} = -2^{-\frac{3}{2}}\psi_0^2$. But $P_{0,0}$ is real, so

$$\psi_2^0 = \overline{\psi}_2^0 = \text{const.} \tag{4·21}$$

Thus, from (4·6),

$$\bar{\eth}^2\sigma^0 = \eth^2\overline{\sigma}^0. \tag{4·22}$$

This is precisely the condition that σ^0 should have no 'magnetic' part and is the condition (since $\dot{\sigma}^0 = 0$) for the existence of a supertranslation which transforms σ^0 to zero everywhere (Newman & Penrose 1966). Hence, without loss of generality we can suppose our coordinates to be chosen so that

$$\sigma^0 = 0 \tag{4·23}$$

everywhere, while retaining the u-independence of all the ψ_m^n, etc. It is important to realize, however, that a supertranslation may be involved in obtaining the condition (4·23). This will be of relevance later, since when we come to consider transitions between stationary states, we shall have to consider the possibility that the same coordinate system cannot be used to cover *both* stationary regions of space-time, while retaining 'manifest stationarity' and condition (4·23). For this reason, the supertranslation invariance of the G_m will be of importance and is derived in § 5.

Now, in addition to (4·20), we have, from (4·10), (4·12), (4·13),

$$\eth\psi_1^0 = 0, \tag{4·24}$$

$$\bar{\eth}\eth\psi_0^0 = 0 \quad \text{whence} \quad \eth\psi_0^0 = 0 \tag{4·25}$$

and

$$\bar{\eth}\eth\psi_0^1 + 6\psi_0^1 = 10\{(\psi_1^0)^2 - \psi_0^0\psi_2^0 - \tfrac{1}{2}\psi_0^0\overline{\psi}_2^0\}. \tag{4·26}$$

Equations (4·24), (4·25) can be integrated immediately by (2·6) and (2·7), and by using the fact that ψ_1^0 and ψ_0^0 are of spin-weight 1 and 2 respectively, to yield

$$\psi_1^0 = \sum_m B_{m\,1}Y_{1,m} \quad (m = -1, 0, 1,); \tag{4·27}$$

$$\psi_0^0 = \sum_m C_{m\,2}Y_{2,m} \quad (m = -2, -1, 0, 1, 2). \tag{4·28}$$

Equation (4·26) can be integrated by first noting that since $\eth$ annihilates the right-hand side of (4·26), this being of spin weight 2, this right-hand side is a sum over ${}_2Y_{2,m}$. Using (2·7), we get

$$\psi_0^1 = \sum_m D_{m\,2}Y_{3,m} + H \tag{4·29}$$

with

$$H = \tfrac{5}{3}(\psi_1^0)^2 - \tfrac{5}{2}\psi_0^0\psi_2^0, \tag{4·30}$$

the first term being the solution to the homogeneous equation. The conserved quantities now are

$$G_m = \int {}_2\overline{Y}_{2,m} H \, d\omega. \tag{4·31}$$

They are in general different from zero. In fact, ψ_2^0, B_m and C_m are respectively proportional to the mass, dipole moment and quadrupole moment as measured by the field at large distances. Thus for any *stationary* (*asymptotically flat*) *gravitational field*, the G_m are explicit quadratic expressions in these first three moments. In particular, in the axially symmetric cases (including the Weyl–Levi–Civita solutions, obtained by making all the arbitrary constants real, cf. Janis & Newman 1965) we have, from (4·27), (4·28), (4·31):

$$\psi_2^0 = -2\sqrt{2}\,M, \quad \psi_1^0 = 6D\sin\theta, \quad \psi_0^0 = -3\sqrt{2}\,Q\sin^2\theta;$$

$$G_0 = 8\sqrt{(30\pi)}\,(2D^2 - MQ), \quad G_{-2} = G_{-1} = G_1 = G_2 = 0.$$

Here M, D and Q are the mass and complex dipole and quadrupole moments.

The non-vanishing of G_m for the general static or stationary case leads to what appears, at first sight, to be a somewhat remarkable state of affairs. Suppose an initially stationary *asymmetric* body, with $H_{\text{initial}} = \frac{5}{3}(\psi_1^0)^2 - \frac{5}{2}\psi_0^0\psi_2^0 \neq 0$, alters its shape to become *spherically symmetric* at a later time. (Actually only the first three moments will need to exhibit this spherical symmetry.) In the meantime the system radiates gravitationally (of necessity, cf. (4·10), (4·9), (4·11), (4·5), (4·4)) and H_{final} will be zero. The $G_m = \int {}_2\overline{Y}_{2,m} H_{\text{initial}}\,d\omega$ must remain unchanged, but because $H_{\text{initial}} \neq H_{\text{final}}$ the final solution can no longer satisfy the conditions which led to (4·31); that is to say, be independent of u with $\sigma^0 = 0$. Assuming the results of § 5, this implies that the final state *cannot be stationary*. For otherwise a Bondi–Metzner–Sachs transformation (cf. Sachs 1962*b*) could be found so that the *new* coordinates would satisfy the above conditions. This might involve a Lorentz velocity transformation (if non-zero resultant momentum is carried by the waves) and perhaps a supertranslation. Thus, from the fact that the G_m transform as $D(2,0)$ under Lorentz transformation and are invariant under supertranslation it follows that the G_m cannot all vanish for the final state in the new coordinate system, contradicting (4·31).

Note that the final field is not spherically symmetrical (e.g. by Birkhoff's theorem or simply because $G_m \neq 0$) even though the source *is* a spherically symmetrical body. Thus the asymmetry in the field must be due to gravitational waves. This disturbance would have to be present at *all* retarded times in the final state, even after the source itself may have ceased emitting waves. The waves would therefore have to possess an effective *incoming* component. In fact, this can be seen more explicitly from (4·13) which now reads (when $H = 0$)

$$\dot{\psi}_0^2 = -\tfrac{1}{2}(\bar{\eth}\eth\psi_0^1 + 6\psi_0^1). \tag{4·32}$$

The $\psi_0^1 = \sum\limits_m G_{m\,2}Y_{2,m} +$ (other terms with different angular behaviour) now yields a linearly time-dependent term in ψ_0^2 proportional to $\Sigma \mathrm{u} G_{m\,2}Y_{2,m}$. This term is identical with the time-dependent term of the linear theory in (3·34), which arose from the incoming radiation field. (In order to remove this time-dependence in the hope of making the system stationary again, the moments would have to be adjusted so that $H_{\text{final}} = H_{\text{initial}}$.) Thus, during the radiative period and afterwards, G_m represents not only information about the multipole structure of the source but in addition, a part apparently corresponding to incoming (i.e. back-scattered) radiation which destroys the stationary nature of the field.

It should be emphasized that although it *appears* as if these solutions (with the linear time dependence in ψ_0^2) are not only non-stationary but perhaps even 'blowing up' with time, this need *not* be the case. It seems highly likely that the back-scattered waves will (in general) *die off* with time so the system should become asymptotically stationary. (If this is true, we can refer to the back-scattered waves as 'tails' to the outgoing waves.† To emphasize the plausibility of this, let us consider an analogous situation for a linear theory. For simplicity we shall consider the spin zero case. The behaviour in linear gravitational theory will be exactly analogous. The example is similar to those of (3·19) and (3·32), but is here non-singular and explicit.

The complex scalar field (Lanczos 1930)

$$\chi = \frac{2a}{(t+\mathrm{i}\sqrt{2}\,b)^2-x^2-y^2-z^2} = \frac{a}{(\mathfrak{u}+ib)(\mathfrak{u}+\mathfrak{r}+ib)} = \frac{a}{\mathfrak{r}(\mathfrak{u}+ib)} - \frac{a}{\mathfrak{r}(\mathfrak{u}+\mathfrak{r}+ib)} \tag{4·33}$$

(a, b constant, with b real) satisfies the D'Alembert equation $\Box\chi = 0$ in flat space-time and is *non-singular* everywhere as is evident from the form with Cartesian x, y, z, t coordinates. It represents an incoming–outgoing wave which dies off in time as t^{-2}, in any fixed region of space $\mathfrak{r} < \mathfrak{r}_0$. Expanding in powers of $\mathfrak{r}^{-1}$, we get

$$\chi = \frac{a}{\mathfrak{u}+ib}\,\mathfrak{r}^{-1} - a\mathfrak{r}^{-2} + a(\mathfrak{u}+ib)\,\mathfrak{r}^{-3} - a(\mathfrak{u}+ib)^2\mathfrak{r}^{-4} + \dots. \tag{4·34}$$

The coefficient $-a$ of $\mathfrak{r}^{-2}$ is effectively the conserved analogue of F_m and G_m for the scalar case (see § 5). Also, similarly to the case of ψ_0^2 above, we also have a linear $\mathfrak{u}$-dependence in the next coefficient, $a(\mathfrak{u}+ib)$, the coefficient of $\mathfrak{u}$ being proportional to the conserved quantity $-a$. It seems not unreasonable to suppose that, in the non-linear gravitational situation considered above, the back-scattered wave dies off with time in a qualitatively somewhat similar way to (4·33) (or, perhaps more appropriately, to a linear gravitational field like that obtained by using χ as a Hertz potential, cf. Penrose 1965). If so, then we can fairly say that the above non-linear field does become asymptotically stationary.

The conserved quantities G_m, which have a quadrupole type structure, can be related to a trace-free symmetric three-dimensional tensor K_{ij} (cf. Jackson 1962) by

$$\left.\begin{aligned} G_2 &= 8\sqrt{(5\pi)}\,(K_{11}-2\mathrm{i}K_{12}-K_{22}),\\ G_1 &= -16\sqrt{(5\pi)}\,(K_{13}-\mathrm{i}K_{23}),\\ G_0 &= 8\sqrt{(30\pi)}\,K_{33},\\ G_{-1} &= -16\sqrt{(5\pi)}\,(K_{13}+\mathrm{i}K_{23}),\\ G_{-2} &= 8\sqrt{(5\pi)}\,(K_{11}+2\mathrm{i}K_{12}-K_{22}).\end{aligned}\right\} \tag{4·35}$$

with $0 = K_{11}+K_{22}+K_{33}$.

The tensor K_{ij}, for *stationary* fields, must be a quadratic function of the dipole moment and bilinear in the monopole and quadrupole moments. In addition (see §§ 1 and 5) it must be invariant under translations of the origin. These conditions plus the requirement that K_{ij} be trace-free leads to

$$K_{ij} = \gamma\{3D_iD_j - \delta_{ij}D_kD_k - MQ_{ij}\}, \tag{4·36}$$

where M, D_i and Q_{ij} are the mass, dipole moment vector and quadrupole moment tensor, and where γ is a numerical constant. By comparing with the axially symmetric case, we see that $\gamma = 1$. For *static* fields, D_i, Q_{ij}, (and K_{ij}) are *real*.

---

† Bonnor & Rotenberg (1966) had previously suggested that the presence of 'tails' was likely to be a general feature of gravitational radiation.

New conservation laws for zero rest-mass fields in asymptotically flat space-time. *Proc. Roy. Soc. Lond. Ser. A* **305(1481)** (1968), 175–204

5. Alternative approach to the conservation laws

In this section the new conservation laws are approached from a different point of view. Although, at first sight, the approach may appear as something of a digression, it enables more insight to be gained into the meanings of the conservation laws, particularly in the linear theories, and into their transformation properties. It also leads to a complete proof of the *supertranslation invariance* of the G_m (Einstein vacuum equations) and of the F_m, N_m (Einstein–Maxwell–neutrino equations)† for asymptotically flat space-times. (A more general and explicit discussion of the transformation behaviour encountered here, and some other related results, will be given elsewhere; cf. also Penrose 1967.)

Let us, for generality, consider the zero rest-mass free-field equations for arbitrary spin $s = 0, \frac{1}{2}, 1, \ldots$. We are interested, here, primarily in the cases $s = 1$ (electromagnetism) and $s = 2$ (gravitation), but the results will apply also to the wave equation ($s = 0$) and the Weyl neutrino equation ($s = \frac{1}{2}$). For $s > 0$, the relevant equation is

$$\nabla^{\mathrm{AP}'}\eta_{\mathrm{AB}\ldots\mathrm{K}} = 0, \tag{5·1}$$

where $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ is a symmetric spinor with $2s$ indices representing the field. (Capital roman indices are two-component spinor indices running over the values 0, 1. Primed indices refer to the complex conjugate spin space. See Penrose (1965) for the notation used here and for other references.) The operator $\nabla^{\mathrm{AP}'}$ is a spinor translation ($\nabla^{\mathrm{AP}'} = \sigma_{\mu}^{\mathrm{AP}'}\nabla^{\mu}$) of the covariant differentiation operator ∇_{μ}.

In flat space-time, (5·1) leads to the 'wave equation'

$$\nabla_{\mu}\nabla^{\mu}\eta_{\mathrm{AB}\ldots\mathrm{K}} = 0, \tag{5·2}$$

but in curved space-time, (5·2) has to be modified by the introduction of additional terms involving the curvature. Also, if $s \geqslant \frac{3}{2}$, (5·1) implies a connexion between $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ and the conformal curvature spinor Ψ_{ABCD}, namely

$$\eta_{\mathrm{ABC(DE}\ldots\mathrm{K}}\Psi^{\mathrm{ABC}}{}_{\mathrm{L})} = 0 \tag{5·3}$$

(round brackets denote symmetrization). This follows because of Ricci identities (see Plebanski 1965; Buchdahl 1958, 1962). However, in the case of gravitation ($s = 2$) we can put $\eta_{\mathrm{ABCD}} = \Psi_{\mathrm{ABCD}}$ so that (5·3) is automatically satisfied, (5·1) being the empty-space Bianchi identity. In the case of the Einstein–Maxwell–neutrino equations, (5·1) (with $s = 1$) becomes the spinor translation of the Maxwell free-field equations and (with $s = \frac{1}{2}$) it becomes the Weyl–neutrino equation $\nabla^{\mathrm{AP}'}\nu_{\mathrm{A}} = 0$. When $s = 0$, we may take for the analogue of (5·1)

$$(\nabla_{\mu}\nabla^{\mu} + \tfrac{1}{6}R)\,\eta = 0, \tag{5·4}$$

where R is the scalar curvature ($R = 0$ in flat space-time) since this has the desired conformal invariance properties.‡

---

† With a notation for the neutrino field similar to that of § 3, we have $N_m = \int {}_{\frac{1}{2}}Y_{\frac{1}{2},m}\,\nu_0^1\,\mathrm{d}\omega$, but see footnote on page 176.

‡ This equation is chosen here for $s = 0$ because it simplifies the discussion, but from the point of view of the new conservation laws it actually makes no essential difference whether or not the $R/6$ term is included. A divergence-free energy-momentum tensor for (5·4) is $\nabla_{\mu}\eta\nabla_{\nu}\eta - \frac{1}{4}g_{\mu\nu}\nabla_{\rho}\eta\nabla^{\rho}\eta - \frac{1}{2}\eta\nabla_{\mu}\nabla_{\nu}\eta + \frac{1}{24}R\eta^2 g_{\mu\nu} - \frac{1}{4}\eta^2 R_{\mu\nu}$. This is symmetric and trace-free, but not positive definite. For $s \geqslant \frac{3}{2}$, no energy-momentum tensor exists at all for (5·1).

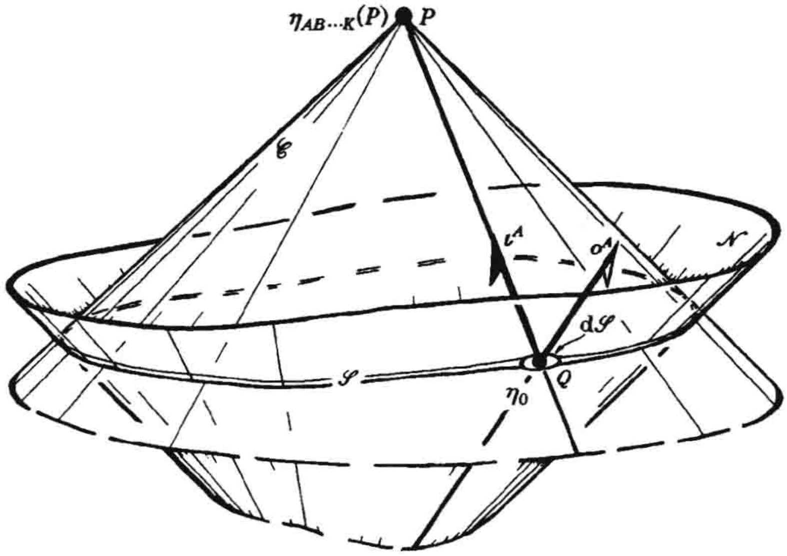


FIGURE 2. The field $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ at P can be expressed in terms of initial data on a null hypersurface $\mathscr{N}$, by means of an integral taken over the intersection of $\mathscr{N}$ with the null cone $\mathscr{C}$ of P.

We shall begin by considering solutions of (5·1) and (5·4) in *flat* space-time. A generalized Kirchoff integral can then be used to express the field at any point P, as an integral of initial data on an initial null hypersurface $\mathscr{N}$. The integral is taken over the two-dimensional intersection, $\mathscr{S} = \mathscr{N} \cap \mathscr{C}$, of $\mathscr{N}$ with the null cone $\mathscr{C}$ of P (figure 2). The required formula (Penrose 1963, 1964) is

$$\eta_{\mathrm{AB}\ldots\mathrm{K}}(P) = \frac{(-1)^{2s}}{2\pi}\int \frac{1}{p}\,\iota_{\mathrm{A}}\iota_{\mathrm{B}}\ldots\iota_{\mathrm{K}}\,\mathscr{D}\eta_0\,\mathrm{d}\mathscr{S} \tag{5·5}$$

integrated over all points $Q \in \mathscr{S}$, with $\mathrm{d}\mathscr{S}$ the surface area element of $\mathscr{S}$ at Q. The *null-datum* η_0 is given by

$$\eta_0 = o^{\mathrm{A}}o^{\mathrm{B}}\ldots o^{\mathrm{K}}\eta_{\mathrm{AB}\ldots\mathrm{K}}, \tag{5·6}$$

o^{A} being a spinor defined on $\mathscr{N}$ which represents the null direction in $\mathscr{N}$. We choose o^{A} normalized at points of $\mathscr{S}$ by

$$o_{\mathrm{A}}\iota^{\mathrm{A}} = 1, \tag{5·7}$$

$p\iota_{\mathrm{A}}\bar{\iota}_{\mathrm{B}'}$ being the spinor translation of the null vector $\overrightarrow{QP}$, so p measures the affine extent of $\overrightarrow{QP}$ (or, what in flat space-time amounts to the same thing, the luminosity distance QP). The operator $\mathscr{D}$ is defined, here, by

$$\mathscr{D} \equiv D - 2s(\rho+\epsilon) - \rho \tag{5·8}$$

where $$D \equiv o^{\mathrm{A}}\bar{o}^{\mathrm{B}'}\nabla_{\mathrm{AB}'}, \quad \rho = o^{\mathrm{A}}\iota^{\mathrm{B}}\bar{o}^{\mathrm{C}'}\nabla_{\mathrm{BC}'}o_{\mathrm{A}}, \quad \epsilon = \iota^{\mathrm{A}}Do_{\mathrm{A}} \tag{5·9}$$

(see Newman & Penrose 1962). Thus D denotes differentiation along the generators of $\mathscr{N}$, and ρ $(=\bar{\rho})$ measures the convergence of these generators. The quantity ϵ is reduced to zero if o^{A} is chosen constant along the generators of $\mathscr{N}$.

The significance of the combination (5·8) is that, in the first instance, under a rescaling

$$o^{A} \rightarrow \chi o^{A}, \quad \iota^{A} \rightarrow \chi^{-1}\iota^{A} \tag{5·10}$$

we have $\mathscr{D}\eta_0 \rightarrow \chi^{2s+1}\bar{\chi}\mathscr{D}\eta_0$. Hence the integral (5·5) is invariant under (5·10) (with $p \rightarrow \chi\bar{\chi}p$). Secondly, under the conformal transformation

$$\tilde{g}_{\mu\nu} = \Omega^{-2}g_{\mu\nu}, \quad \tilde{\sigma}_{\mu}^{AB'} = \Omega^{-1}\sigma_{\mu}^{AB'} \tag{5·11}$$

we have (see Penrose 1965)

$$\tilde{\mathscr{D}}(\Omega\xi) = \Omega\mathscr{D}\xi, \tag{5·12}$$

where we choose

$$\tilde{o}^{A} = \Omega^{-\frac{1}{2}}o^{A}, \quad \tilde{\iota}^{A} = \Omega^{\frac{1}{2}}\iota^{A}, \tag{5·13}$$

so that

$$\tilde{\rho} = \rho + \Omega^{-1}D\Omega, \quad \tilde{\epsilon} = \epsilon - \Omega^{-1}D\Omega. \tag{5·14}$$

In order to preserve the field equations (5·1), (5·4) under (5·11), we define the conformally transformed field by

$$\tilde{\eta}_{AB\ldots K} = \Omega^{s+1}\eta_{AB\ldots K} \tag{5·15}$$

(see Penrose 1965). Thus $\tilde{\eta}_0 = \Omega\eta_0$, by (5·6), (5·13), whence by (5·12) we have

$$\tilde{\mathscr{D}}\tilde{\eta}_0 = \Omega\mathscr{D}\eta_0$$

so $\mathscr{D}$ is a conformally invariant differentiation operator for η_0 in the null direction on $\mathscr{N}$.

Although we cannot, in general, expect (5·5) always to remain true in *curved* space-time, we may still *perform* the integral in (5·5) in curved space-time provided we interpret $\iota_A\iota_B\ldots\iota_K$ as being in the tangent space at P rather than at Q. For each choice of Q we must refer the ι_A at Q to the point P by parallel transport along the null geodesic QP. The choice of conformal transformation behaviour for ι_A, given in (5·13), is made so that the property of ι_A being transported parallelly along QP is preserved under (5·11). We now define the measure of extent, p, of the geodesic QP to be the *luminosity* distance from Q to P scaled in terms of the vector $n^{\mu} = \sigma^{\mu AB'}\iota_A\bar{\iota}_B$. That is to say, $p^2\,\mathrm{d}\Sigma = \mathrm{d}\mathscr{S}$ where $\mathrm{d}\Sigma$ is the element of area traced out in the tangent space at P by the vector n^{μ}. (In flat space-time this agrees with the previous definition of p.)

Under the conformal transformation (5·11) we then have

$$\tilde{p} = \overset{P}{\Omega}{}^{-1}\overset{Q}{\Omega}{}^{-1}p, \tag{5·16}$$

where the upper indices P, Q denote different points to which the conformal factor Ω refers. Under (5·11) the remaining quantities involved in the integral (5·5) transform as

$$\iota_A = \overset{P}{\Omega}{}^{\frac{1}{2}}\iota_A, \quad \tilde{\mathscr{D}}\tilde{\eta}_0 = \overset{Q}{\Omega}\mathscr{D}\eta_0, \quad \widetilde{\mathrm{d}\mathscr{S}} = \overset{Q}{\Omega}{}^{-2}\,\mathrm{d}\mathscr{S} \tag{5·17}$$

whence the result of the integration is conformally covariant, transforming according to (5·15). The validity of (5·5) for a particular choice of metric therefore implies its validity for any conformally equivalent metric. Hence, its validity in any *conformally flat* space-time follows from its validity in Minkowski space-time. One proof of (5·5) which applies in flat and hence, in conformally flat space-time has been given earlier (Penrose 1963). In fact, (5·5) is in general *not* valid in *conformally curved* space-times.) Another way of obtaining the result will be implicit in the following arguments.

The significance of the conformal invariance of (5·5) for the present paper lies in the fact that in flat and asymptotically flat space-times (defined appropriately) it is possible to transform the metric of the space-time so that the past and future null *infinities* become *null cones* $\mathscr{I}^-$, $\mathscr{I}^+$, respectively, (which constitute a boundary to the transformed manifold $\mathscr{M}$) by means of conformal transformations (Penrose 1965). By allowing $\mathscr{I}^+$, say, to take the place of $\mathscr{C}$ in (5·5), we obtain a statement about asymptotic properties of the field $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ in the untransformed (i.e. physical) space-time. For it turns out that if the appropriate conditions (which amount, in effect, to (4·1), (4·2), together with (3·10), etc.; see Penrose 1965) are imposed at infinity, then the space-time is sufficiently conformally *flat* near $\mathscr{I}^+$ to enable (5·5) to be effectively valid with $\mathscr{C} = \mathscr{I}^+$.

However, except in certain cases for which all the sources are finally radiated away in the form of zero rest-mass radiation, it turns out, in general relativity, that the vertex $P = I^+$ (and also I^0, see Penrose 1965) of $\mathscr{I}^+$ is a singular point of $\mathscr{M}$, while in the case of linear theories, the free-field equation (5·1) or (5·4) will generally be violated at I^+. Thus, we cannot usually expect to employ (5·5) just as it stands. We can, however, simply use the fact that the integral in (5·5) yields the same result whichever cross-section $\mathscr{S}$ of $\mathscr{C}$ is used. Indeed, whether $\mathscr{C}$ represents $\mathscr{I}^+$ or a finite cone, we can evaluate the integral in (5·5) whenever the field $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ is smoothly defined on $\mathscr{S}$. This will give us a *formal* value of the 'field at P' (which would agree with the actual value of the field at P whenever the field equation (5·1) or (5·4) is satisfied in the neighbourhood of the entire cone $\mathscr{C}$—or that portion of $\mathscr{C}$ connecting

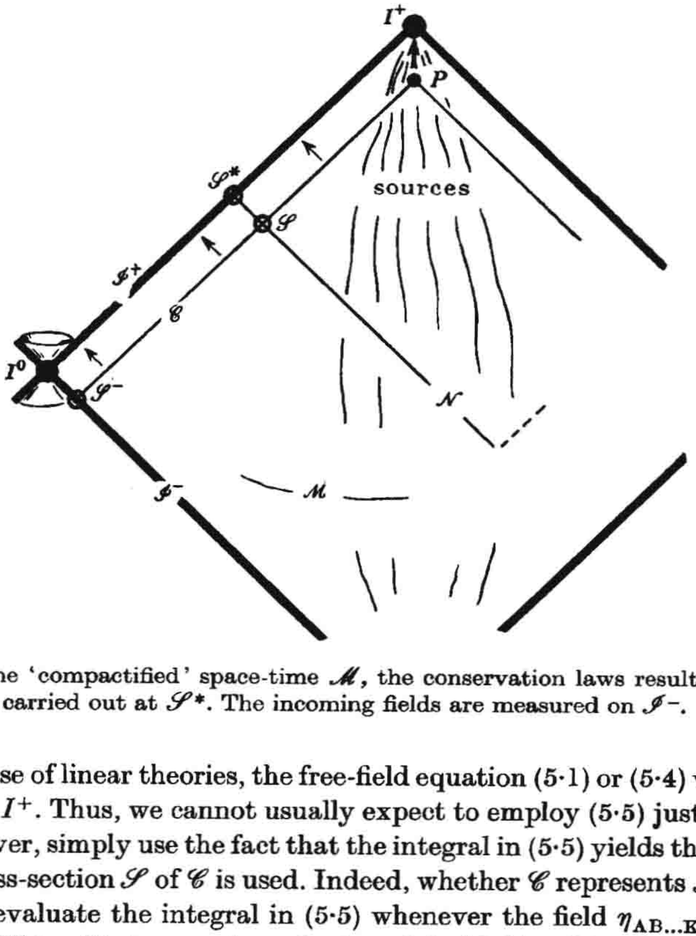


FIGURE 3. In the 'compactified' space-time $\mathscr{M}$, the conservation laws result from integrals carried out at $\mathscr{S}^*$. The incoming fields are measured on $\mathscr{I}^-$.

$\mathscr{S}$ to P). This will work even when $P = I^+$ is a singular point of the conformal manifold $\mathscr{M}$, and we should expect this formal 'field at I^+' to be unchanged as the cross-section of $\mathscr{I}^+$ is varied, provided that (5·1) or (5·4) is satisfied in the neighbourhood of that portion of $\mathscr{I}^+$ over which $\mathscr{S}$ moves.

Assuming, for the moment, the validity of this, we shall have an *asymptotic conservation law* for any spin s zero rest-mass field $\eta_{\mathrm{AB}\ldots\mathrm{K}}$, which holds so long as the field equation (5·1) or (5·4) is satisfied near infinity. Indeed, since $\eta_{\mathrm{AB}\ldots\mathrm{K}}$ has $(2s+1)$ formal 'components at I^+', we should get a set of $(2s+1)$ conserved quantities for each spin s field (complex, if $s > 0$). Each cross-section $\mathscr{S}$ of $\mathscr{I}^+$ defines the points at infinity of a definite null hypersurface $\mathscr{N}$. (Generate $\mathscr{N}$ by the null geodesics, other than those lying on $\mathscr{I}^+$, which meet $\mathscr{S}$ orthogonally.) Then, in the untransformed physical space-time, $\mathscr{S}$ represents the points at future null infinity of an outgoing null hypersurface $\mathscr{N}$. The surface independence of the integral on $\mathscr{I}^+$ thus expresses the independence of the integral, of the particular choice of null hypersurface used to define it, i.e. it expresses the *supertranslation invariance* of the conserved quantities.

To see that the above definition of the conserved quantities in fact agrees with the definitions (3·17), (3·31) given in § 3, we must examine the meaning of the integral in (5·5) when $\mathscr{C}(=\mathscr{I}^+)$ represents future null infinity. If the null-datum η_0 (when conformally transformed so that $\mathscr{I}^+$ is finite) is C^N *differentiable* at $\mathscr{S}$, then, in the untransformed physical space-time, we have

$$\eta_0 = \sum_{j=0}^{N} \eta_0^j \mathfrak{r}^{-2s-1-j} + \mathrm{o}(\mathfrak{r}^{-2s-1-N})$$

where, along each generator of $\mathscr{N}$, the $\eta_0^0, \eta_0^1, \ldots, \eta_0^N$ are constants and $\mathfrak{r}$ is a (physical) affine or luminosity parameter. The (conformally transformed) value of η_0 on $\mathscr{S}$ is then given by the leading coefficient η_0^0 (see Penrose 1965). Furthermore, the value of $\mathscr{D}\eta_0$ on $\mathscr{S}$ is given by the second coefficient η_0^1, as is not hard to verify. The remaining coefficients $\eta_0^2, \ldots, \eta_0^N$ give the higher derivatives $\mathscr{D}^2\eta_0, \ldots, \mathscr{D}^N\eta_0$ of η_0 at $\mathscr{S}$ in effect, but possibly modified by terms involving lower derivatives. (This depends on the precise choice of parameter and conformal factor).

Thus, it is η_0^1 which appears in the integrand in (5·5) when $\mathscr{C} = \mathscr{I}^+$. The quantity p becomes effectively a constant and can be ignored if the scaling of $\mathfrak{r}$ is chosen correctly. (The scaling of $\mathfrak{r}$ is related to the scaling of p because $\mathfrak{r}$ depends on o^{A} and p on ι^{A}, related by (5·7).) The $2s+1$ different components of $\iota_{\mathrm{A}}\iota_{\mathrm{B}}\ldots\iota_{\mathrm{K}}$ define (up to simple factors) the ${}_s\overline{Y}_{s,m}$ for $m = -s, -s+1, \ldots, s$ (see Newman & Penrose 1966, eq. (3·21)); the components of ι_{A} can be given as $(1+\zeta\bar{\zeta})^{-s}(1, \zeta)$ in a $(\zeta, \bar{\zeta})$-system of coordinates for the sphere of cross-section of $\mathscr{I}^+$). Omitting a numerical factor for each m, the integral in (5·5) becomes, when $\mathscr{C} = \mathscr{I}^+$:

$$\int {}_s\overline{Y}_{s,m}\,\eta_0^1\,\mathrm{d}\omega \quad (m = -s, -s+1, \ldots, s), \tag{5·18}$$

so that the $2s+1$ quantities (5·18) will be conserved for each m. They will be invariant under supertranslations of the Bondi–Metzner–Sachs group, provided (5·1) or (5·4) holds near $\mathscr{I}^+$. Also, since the formal field '$\eta_{\mathrm{AB}\ldots\mathrm{K}}(I^+)$' is an irreducible spinor of valence $[2s, 0]$ 'at I^+', it follows that (5·18) transform according to a $D(s, 0)$ representation of the Lorentz group under asymptotic Lorentz rotations. When $s = 1$ we get the F_m of (3·17) and when $s = 2$ we get the G_m of (4·19); $s = \frac{1}{2}$ yields the conserved quantities N_m for the ('classical') neutrino field.

It remains to demonstrate the truth of the assertion, made above, that the integral (5·5) is, in fact, independent of the choice of surface $\mathscr{S}$, when $\mathscr{C} = \mathscr{I}^+$. The essential property of $\mathscr{I}^+$ that is required is that it, and its neighbourhood in the (transformed) space-time manifold, are both 'sufficiently conformally flat'. To be more precise, we require that, in the neighbourhood of each point Q of $\mathscr{I}^+$, a choice of conformal factor (and scaling (5·10)) exists for which (according to the transformed metric)

$$\nabla_{AB'}\iota_C = 0 \quad \text{at} \quad \mathscr{I}^+, \tag{5·19}$$

where also

$$C_{\mu\nu\rho\sigma} = 0 \quad \text{at} \quad \mathscr{I}^+. \tag{5·20}$$

In fact (5·20) is effectively equivalent to the peeling-off property (4·1), (4·2) (with $N = 1$) for the gravitational field while the possibility of arranging (5·19) is a consequence of the corresponding peeling-off properties (3·10), etc., for the Maxwell, neutrino and scalar wave fields (see Penrose 1965). In the case of linear fields in flat space-time, we can arrange (5·19) and (5·20) to hold everywhere throughout a *finite* (four-dimensional) neighbourhood in $\mathscr{M}$, of any point of $\mathscr{I}^+$.

We shall employ a form of Gauss–Stokes theorem on $\mathscr{I}^+$. If β_{AB} is any symmetric spinor defined on $\mathscr{I}^+$, then $b_{\mu\nu} = \beta_{AB}\,\sigma_{\mu}{}^{A}{}_{C'}\,\sigma_{\nu}^{BC'}$ is self-dual skew-symmetric bivector on $\mathscr{I}^+$. Hence $\int b_{\mu\nu}\,\mathrm{d}\mathscr{S}^{\mu\nu}$ is independent of the surface of integration, provided $\nabla_{[\lambda}b_{\mu\nu]}\mathrm{d}\mathscr{I}^{\lambda\mu\nu} = 0$, $\mathrm{d}\mathscr{S}^{\mu\nu}$ being the bivector surface element associated with $\mathrm{d}\mathscr{S}$, and $\mathrm{d}\mathscr{I}^{\lambda\mu\nu}$ the trivector 'volume' element of $\mathscr{I}^+$. Since $b_{\mu\nu}$ is self-dual, this condition is equivalent to $n_\mu \nabla_\nu b^{\mu\nu} = 0$, where $n_\mu = \sigma_{\mu}^{AB'}\,\iota_A\,\bar{\iota}_B$ is normal to $\mathscr{I}^+$, i.e. the condition is equivalent to

$$\bar{\iota}_{C'}\,\iota^B\,\nabla^{AC'}\beta_{AB} = 0. \tag{5·21}$$

Thus (5·21) implies that

$$\int \beta_{AB}\,o^A\,\iota^B\,\mathrm{d}\mathscr{S} \tag{5·22}$$

is independent of cross-section $\mathscr{S}$ on $\mathscr{I}^+$, since $\sigma_{\mu}{}^{A}{}_{C'}\,\sigma_{\nu}^{BC'}\,\mathrm{d}\mathscr{S}^{\mu\nu} = 2\mathrm{i}o^{(A}\iota^{B)}\,\mathrm{d}\mathscr{S}$. If we put

$$\alpha_A = \beta_{AB}\,\iota^B \tag{5·23}$$

and employ (5·19), we get the result that if

$$\bar{\iota}_{C'}\,\nabla^{AC'}\,\alpha_A = 0, \tag{5·24}$$

then the value of

$$\int \alpha_A\,o^A\,\mathrm{d}\mathscr{S} \tag{5·25}$$

is unchanged as $\mathscr{S}$ varies over any region of $\mathscr{I}^+$ for which (5·19) also remains true. (Given α_A, we can always find a symmetric β_{AB} satisfying (5·23), for example, $\beta_{AB} = (\alpha_C\,\iota^C)^{-1}\alpha_A\,\alpha_B$.)

We are now in a position to outline the verification of the surface independence of (5·5), (5·18). Put

$$\bar{\alpha}_{\mathrm{M}} = o^{\mathrm{A}} o^{\mathrm{B}} \dots o^{\mathrm{K}} \{ o^{\mathrm{L}} \nabla_{\mathrm{LM'}} \eta_{\mathrm{AB}\dots\mathrm{K}} + (2s+1)\, \eta_{\mathrm{AB}\dots\mathrm{K}} \nabla_{\mathrm{LM'}} o^{\mathrm{L}} \}. \tag{5·26}$$

Then it follows, with the aid of (5·7), (5·19), (5·20) and (5·1) or (5·4), that (5·24) holds. Furthermore, by (5·6), (5·7), (5·8), (5·9) and (5·19),

$$\bar{\alpha}_{\mathrm{M'}} \bar{o}^{\mathrm{M'}} = \overline{\mathscr{D}\eta_0}$$

whence it follows from (5·25) that $\int \mathscr{D}\eta_0 \, \mathrm{d}\mathscr{S}$ does not change as $\mathscr{S}$ is varied over any finite region of $\mathscr{I}^+$ for which (5·19) remains true. But this integral is, with the different scaling of ι_{A} chosen here for (5·19), effectively one component of (5·18). By choosing several conformal factors and scalings for ι_{A} consistent with (5·19) we can generate a set of quantities on which *all* components of (5·18) are linearly dependent. Hence, (5·18) remains constant as $\mathscr{S}$ varies over any 'finite' region of $\mathscr{I}^+$. Consequently, the value of (5·18) is independent of the cross-section $\mathscr{S}$ of $\mathscr{I}^+$, as required. (A more explicit discussion of the transformation expressions involved here is to be given elsewhere.) Thus, in particular, the quantities F_m of (3·17) in the Einstein–Maxwell (–neutrino) theory and the G_m of (4·19) in empty-space Einstein theory are both *supertranslation invariant.*

In the case of the linear theories in flat space-time the arguments can be carried somewhat further. In the first instance, since (5·19) and (5·20) can be made to hold throughout a *finite* neighbourhood, in $\mathscr{M}$, of any point of $\mathscr{I}^+$, it follows that the *derivatives* of (5·19) and (5·20) will also hold at $\mathscr{I}^+$. This implies that we can use not only the integral in (5·5) when $\mathscr{C} = \mathscr{I}^+$, but also its various derivatives in different directions. Effectively, this means we can employ the formulae analogous to (5·5), which express the various derivatives

$$\nabla_{\mathrm{LQ'}} \eta_{\mathrm{AB}\dots\mathrm{K}}(P), \quad \nabla_{\mathrm{LQ'}} \nabla_{\mathrm{MR'}} \eta_{\mathrm{AB}\dots\mathrm{K}}(P), \dots, \tag{5·27}$$

of $\eta_{\mathrm{AB}\dots\mathrm{K}}$ at P, in terms of the null-datum η_0 on $\mathscr{N}$ (Penrose 1963). Translating the surface independence of these integrals, when $P = I^+$, into physical space terms (i.e. in terms of the η_0^j) we obtain the conservation of the higher order quantities $F_m^{n,0}$, $G_m^{n,0}$ (see (3·17), (3·31)) together with the time-dependence (3·18), (3·30) of the $F_m^{n,l}$ and $G_m^{n,l}$. These respectively form a $D(\frac{1}{2}n+1, \frac{1}{2}n)$ and a $D(\frac{1}{2}n+2, \frac{1}{2}n)$ representation of the Lorentz group ($D(\frac{1}{2}n+s, \frac{1}{2}n)$ for general spin s).

The reason for the more complicated $\mathfrak{u}$-dependence (cf. (3·18), (3·30)) for the $F_m^{n,l}$, $G_m^{n,l}$ when $l \neq 0$ is that although the integrals for the quantities (5·27) are still surface independent, the differential operator $\nabla_{\mathrm{LQ'}}$ is not conformally covariant. If some different conformal factor were chosen, the definition of each of the quantities (5·27) would be modified by the addition of terms in lower order derivatives of $\eta_{\mathrm{AB}\dots\mathrm{K}}$ at P. The interpretation of the integrals for (5·27) which leads to the $F_m^{n,l}$, $G_m^{n,l}$, etc., requires a particular choice of conformal factor which is tied to the definition of $\mathfrak{u}$. Thus, the quantities arising from (5·27) which are conserved in a stronger sense, namely time- *and* space-translation invariance, must involve $\mathfrak{u}$ *explicitly* (i.e. the conserved quantities would have to be polynomials of degree n in $\mathfrak{u}$). This would be an equivalent way of stating (3·18), (3·30).

The conformal technique also affords a direct way of illustrating the connexion between $F_m^{n,l}$, $G_m^{n,l}$ and the incoming radiation fields. Let P be any point in the 'compactified' *Minowski* space-time $\mathscr{M}$—with *boundary* $\dot{\mathscr{M}} = I^- \cup \mathscr{I}^- \cup I^0 \cup \mathscr{I}^+ \cup I^+$. (Any null geodesic, not on $\mathscr{I}^\pm$, will have a past end-point on $\mathscr{I}^-$ and a future end-point on $\mathscr{I}^+$.) Let $\mathscr{C}$ be the past null cone of P (figure 3). We consider the limit $P \to I^+$, so that $\mathscr{C} \to \mathscr{I}^+ \cup I^0 \cup I^+$. Let $\mathscr{N}$ be an outgoing null hypersurface in $\mathscr{M}$ and let $\mathscr{S} = \mathscr{N} \cap \mathscr{C}$. As $P \to I^+$, we have $\mathscr{S} \to \mathscr{S}^* = \mathscr{N} \cap \mathscr{I}^+$. The $F_m^{n,l}$, $G_m^{n,l}$ can be calculated as integrals over the surface $\mathscr{S}^*$. (Strictly, when $n > 0$, we have only considered, in § 3, the case when $\mathscr{N}$ is a cone.) Now, from the continuity at $\mathscr{I}^+$ of the quantities involved, we can regard the $F_m^{n,l}$, $G_m^{n,l}$ as limits, as $P \to I^+$, of the corresponding integrals over $\mathscr{S}$. These integrals are (5·5) and the corresponding expressions which, in the absence of sources, yield the differentiated (formal) fields (5·27) at P.

We must bear in mind that there may be sources to the fields present in the region of $\mathscr{C}$ which connects P to $\mathscr{S}$—so that the results of the integrals refer, not to the actual fields in the neighbourhood of P, but to the 'background' fields in the neighbourhood of P, i.e. to the fields with the retarded contributions to the fields from these sources subtracted out. We shall suppose, however, that P is close enough to I^+ that there are no sources between $\mathscr{S}$ and $\mathscr{S}^- = \mathscr{C} \cap \mathscr{I}^-$ (the sources being 'isolated'). The 'background' fields near P then result solely from the *incoming* radiation fields. For, the integrals at $\mathscr{S}$ yield the same result as the integrals at $\mathscr{S}^-$ (the free field equations (5·1) holding in the region of $\mathscr{C}$ which connects $\mathscr{S}^-$ to $\mathscr{S}$) and the integrals at $\mathscr{S}^-$ are expressed solely in terms of the null-data on $\mathscr{I}^-$, namely the *incoming radiation fields* (Penrose 1965).

The $F_m^{n,l}$, $G_m^{n,l}$ are obtained as $P \to I^+$. But the limit of $\mathscr{S}^-$ is the point I^0. Thus the limiting values of the integrals describe the null data on $\mathscr{I}^-$ in the immediate neighbourhood of the point I^0. In more physical terms, this is the ultimate limiting behaviour of the incoming fields. In fact, the $F_m^{n,l}$, $G_m^{n,l}$ can be used as coefficients in a power series expansion of the incoming fields, by developing the null-data on $\mathscr{I}^-$ as a Taylor series centred on the point I^0. (This is essentially the method developed in Penrose 1963, as applied to the point I^0, and will be discussed in more detail elsewhere.)

When the space-time is not flat, but only asymptotically flat, we might expect to carry over some of these ideas. However, the crucial point whereby the integrals at $\mathscr{S}$ can be equated to integrals at $\mathscr{S}^-$ now fails owing to the failure of the conformal flatness conditions on $\mathscr{C}$. We could try to move the hypersurface $\mathscr{N}$ near to $\mathscr{I}^-$ so that the integrals at $\mathscr{S}$ differ but little from the integrals at $\mathscr{S}^-$ and use the constancy of the F_m, G_m under change of $\mathscr{S}^*$ to try to get an identification of the F_m, G_m in terms of incoming fields. However, here two limiting processes are now involved, namely $\mathscr{S} \to \mathscr{S}^-$ and $\mathscr{S} \to \mathscr{S}^*$ and the order in which these are carried out is all-important. This is because I^0 is now a singular point for $\mathscr{M}$. The detailed structure of this singular point thus appears to contain essential information about the multipole structure of the gravitational sources.

6. Discussion

We here briefly summarize some of the pertinent results of the earlier sections. There is, in the full non-linear theory of pure gravitation (with specific boundary conditions to ensure asymptotic flatness) a set of 5 complex quantities G_m which are absolutely conserved (in the sense of § 1). These quantities have a quadrupole structure and transform as a $D(2,0)$ representation of the homogeneous Lorentz group; they are invariant under supertranslations and, in particular, under the four space-time translations. Incoming gravitational radiation which dies off suitably must contribute to the values of these quantities (inferred from the linear theory); in addition there must be a contribution from the first three multipole moments (since for stationary metrics, the G_m are fully equivalent to a combination of these moments). Owing to G_m conservation we discover that certain gravitational states are inaccessible from other ones in a finite time. For example, since the Schwarzschild metric possesses $G_m = 0$ it cannot be reached from the general Weyl–Levi-Civita solutions ($G_m \neq 0$) in a finite time via the emission of gravitational waves. That is to say, such waves cannot be clean cut and a residual disturbance, resembling an incoming wave which presumably normally dies off with time, is necessarily always present within the space-time cone of radiation (except possibly if the moment combination eventually returns to its original value).

In the coupled Einstein–Maxwell theory, the five complex G_m remain absolutely conserved only if modified by the addition of extra terms (discussed elsewhere) involving the electromagnetic field. However, in addition, the three complex F_m of (3·17) are absolutely conserved, being invariant under supertranslations and transforming as a $D(1,0)$ representation of the Lorentz group. If ('classical') neutrino field is also present, two complex supertranslation invariant quantities N_m exist in addition to the F_m, which transform as $D(\frac{1}{2},0)$ under the Lorentz group.

The question that immediately arises (unfortunately without an answer yet) is whether there is anything of fundamental significance or value in these conservation laws. One might speculate as to whether they are in any way derivable from Noether's theorems;† whether there is, in general relativity, an invariance group (in addition to the coordinate group) which could give rise to these new conservation laws. If such is to be found, this would presumably shed some light on the significance of G_m. The asymmetrical time-behaviour of the G_m and F_m is, in this context, somewhat mysterious. Also, there is no *obvious* way of obtaining the G_m and F_m from the Bondi–Metzner–Sachs group, for example.

We could further ask whether the fact that F_m and G_m are absolutely conserved puts them, in any sense, on a par with electric charge or with mass in (linear) gravitation theory in that a meaningful intrinsic absolute invariant can be assigned to an isolated system. For example, we might consider a 'gedanken' situation in which physicists, unfamiliar with the concept of charge, were able only to examine electromagnetic *field* and to do this only in source-free regions of space. They could discover the Maxwell free-field equations and then derive, theoretically, the law $\int \mathbf{E}\,.\,\mathrm{d}\mathbf{S} = \text{const.}$ for any closed surface S surrounding a 'forbidden' region. They could then theoretically

CHAPTER 35

---

† Some interesting suggestions have been put forward by Goldberg (1967) in this connexion.

infer the existence of *electric charge* for such regions, ruling out magnetic charge only from the experimental fact $\int \mathbf{B}.d\mathbf{S} = 0$. A similar (but even less likely) situation would be physicists understanding linear gravitation theory without having observed mass directly. They could correspondingly infer, theoretically, the existence of mass, momentum and angular momentum, from the surface independence of integrals involving only the (asymptotic) field. The situation with the F_m and G_m, *appears* to be quite similar to the above (and, like charge, mass, momentum and spin, the F_m and G_m are independent of choice of origin). Can we then analogously infer the existence of 'generalized charges' for the sources? Unfortunately there is a complication in the case of F_m and G_m, in that the *free-field itself* contributes (time-asymmetrically) via the asymptotic *incoming* field. The status of the analogy is therefore somewhat obscured. More work has to be done to determine precise relationships between F_m, G_m and the sources.

The role played by F_m and G_m conservation in actual physical processes is also difficult to assess. We may speculate as to whether G_m conservation has any real significance in, for example, the late stages of gravitational collapse. In the late stages of asymmetrical collapse, it seems that large reductions in the mass of the material system may occur quite rapidly through gravitational radiation (if calculations based on the linear theory are to be believed). Since the (positive) energy residing in the gravitational waves is lost to the system (owing to the difficulty of reabsorption), we have, in effect, that mass-energy conservation is no longer operative. Thus it may be that the comparative importance of G_m conservation is increased, since G_m is *not* carried away by the waves. It is just conceivable that the behaviour of a collapsing system may be severely restricted by the G_m conservation law, but much more work would have to be done to decide if this is really the case.

To sum up what we may refer to as the 'best available interpretation' of the physical significance of the G_m (and F_m), we may suppose that at any constant retarded time they may be thought of as composed of two parts, one of which is defined by the multipole structure of the source, the other by what appears, at that retarded time, as asymptotic incoming field. Under normal circumstances one has the freedom, by manipulating the sources, to transfer at will, the contribution to the G_m (or F_m) from the source terms to the wave terms and vice versa. But it is a matter of speculation whether this freedom will persist when it is the highly non-linear dynamics of strong gravitational fields which is dominating the source behaviour.

REFERENCES

Bondi, H. 1960 *Nature, Lond.* **186**, 535.
Bondi, H., van der Burg, M. G. J. & Metzner, A. W. K. 1962 *Proc. Roy. Soc.* A **269**, 21.
Bonnor, W. B. 1959 *Phil. Trans.* **251**, 233.
Bonnor, W. B. & Rotenberg, M. A. 1966 *Proc. Roy. Soc.* A **289**, 247.
Buchdahl, H. A. 1958 *Nuovo Cim.* **10**, 96.
Buchdahl, H. A. 1962 *Nuovo Cim.* **25**, 486.
Cornish, F. H. J. 1965 *Proc. Roy. Soc.* A **286**, 270.
Crampin, M. & Foster, J. 1966 *Proc. Camb. Phil. Soc.* **62**, 269.
Exton, A. R. 1967 Conserved quantities in the Einstein–Maxwell–Neutrino theory. Ph.D. Thesis, University of Pittsburgh.

Exton, A. R., Newman, E. T. & Penrose, R. 1968 *J. Math. Phys.* (to appear).
Goldberg, J. N. 1967 Preprint.
Goldberg, J. N., Macfarlane, A. J., Rohrlich, F., Sudarshan, E. C. G. & Newman, E. T. 1968 *J. Math. Phys.* (to appear).
Jackson, J. D. 1962 *Classical electrodynamics*. New York, London: Wiley.
Janis, A. & Newman, E. 1965 *J. Math. Phys.* **6**, 902.
Lamb, D. J. 1966 *J. Math. Phys.* **7**, 458.
Lanczos, C. 1930 *Z. Phys.* **59**, 514.
Møller, C. 1961 *Mat. Phys. Skr. Dan. Selsk.* **1** no. 10.
Newman, E. T. & Penrose, R. 1962 *J. Math. Phys.* **3**, 566; **4**, 998.
Newman, E. T. & Penrose, R. 1965 *Phys. Rev. Lett.* **15**, 231.
Newman, E. T. & Penrose, R. 1966 *J. Math. Phys.* **7**, 863.
Newman, E. T. & Unti, T. W. J. 1962 *J. Math. Phys.* **3**, 891.
Papapetrou, A. 1957 *Ann. Physik.* **20**, 399.
Penrose, R. 1963 In P. G. Bergmann's *A.R.L. Tech. Docum. Rep.* no. 63-56, *Quantization of generally covariant field theories.*
Penrose, R. 1964 Article in *Relativity, groups and topology: the* 1963 *Les Houches lectures.* New York: Gordon and Breach.
Penrose, R. 1965 *Proc. Roy. Soc.* A **284**, 159.
Penrose, R. 1967 In *Relativity theory and astrophysics*, ed. J. Ehlers (American Mathematical Society).
Plebanski, J. 1965 *Acta Phys. Polon.* **27**, 361.
Sachs, R. 1962*a* *Proc. Roy. Soc.* A **270**, 103.
Sachs, R. 1962*b* *Phys. Rev.* **128**, 2851.
Sachs, R. & Bergmann, P. G. 1958 *Phys. Rev.* **112**, 674.
Statchel, J. J. 1966 *J. Math. Phys.* **7**, 1321.
van der Burg, M. G. J. 1966 *Proc. Roy. Soc.* A **294**, 112.
Winicour, J. 1967 Preprint: Some total invariants of asymptotically flat space-times.

36

STRUCTURE OF SPACE-TIME

Battelle Rencontres: 1967 Lectures in mathematics and physics. Ed. C. M. DeWitt and J. A. Wheeler, pp. 121–235. Benjamin, New York, 1968

This write-up of the twelve 'Battelle Lectures' that I gave in the summer of 1967 was what interrupted the completion of Chapter 31. These *Battelle Rencontres* were organized by John A. Wheeler and Cécile M. DeWitt as a means of bringing together various distinguished mathematicians and physicists. It was an insightful thing to do, because of the paucity of cross-fertilization between these two groups for several decades—and things began to change in the years afterwards. Since my interests straddled mathematics and physics, I was asked to give this 12-lecture course, but for several lectures it received a mixed reception. The reason was that I had decided to test the audience with publicly presenting my 'abstract-index notation' for the first time. The mathematicians hated it because they disliked the physicists' 'summation convention', which they thought I was using, and the physicists couldn't see the point of all this 'abstract stuff'. I had been lulled into a false anticipation of harmony because on the way to Seattle I had stopped off at Chicago at the invitation of Saunders Mac Lane, who was very enthusiastic about abstract indices, and he gave a whole seminar about it to the mathematicians at Chicago University.

VII

Structure of Space-Time*

R. PENROSE

1 Introduction *121*
2 The Nature of General Relativity *125*
3 The Abstract Index Notation *135*
4 Space-Times with Spinor Structure *141*
5 The Interpretation of a Spin-Vector *150*
6 Explicit Curvature Formulas *156*
7 Einstein's Equations and Focusing *160*
8 Conformal Infinity *171*
9 Horizons *186*
10 Gravitational Collapse *202*
11 Singularities in Cosmology *223*
References *232*

1 INTRODUCTION

According to present-day theory, all the phenomena of physics take place within the framework of a certain differentiable manifold referred to as the *space-time continuum.* Our familiarity with this idea is such that it is now regarded as almost "obvious" that space and time should constitute such a structure. However, before discussing the nature of this structure, it is worth examining something of what lies behind this belief. Indeed, there is the definite possibility that some future theory may be found which describes nature more accurately than present theory, but for which the differentiable manifold picture of space-time would not be appropriate. We should not close our minds to such a possibility, but also we should keep in mind the extraordinary range over which the present-day view is such an excellent approximation.

---

* Section 10 (and also some small portions of other sections, notably 2 and 9) is based largely on the author's 1966 Adams Prize essay *An Analysis of the Structure of Space-Time.*

The very accurately "locally Euclidean" nature of space, and the continuity of time, would, indeed, seem to have supplied the prime motivation, in the first instance, for the rigorous development of the continuum concept. At the time of Zeno, no such rigorous concept of continuum existed, so that the idea of a limit, in space or time, seemed puzzling. It does not seem puzzling to us today, but perhaps we are wrong not to be puzzled! The standard resolution of Zeno's paradoxes refers more to the *mathematical* continuum concept than to the nature of space-time itself. The view of space-time as forming a continuum would imply that a continuous nature would persist, no matter how much a system is magnified. But it is not at all clear that continuous descriptions are really appropriate on a scale small enough that quantum phenomena become important. For example, at a scale of 10^{-13} cm (approximately the radius of an elementary particle), the mere attempt at localization of the position of a particle to that accuracy will, as a consequence of the uncertainty principle, imply the probable occurrence of a very large momentum, with the implication that *new* particles are created, some of which may be indistinguishable from the original particle. Thus the concept of "position" for the original particle becomes obscured [22]. More alarming, moreover, is the picture presented if we allow ourselves to discuss phenomena at a dimension of the order of 10^{-33} cm. At such a dimension, the quantum fluctuations in the curvature of space-time (if both present-day quantum theory and gravitation theory can be accurately extrapolated to this degree) would be large enough to produce alterations in topology. Thus, the view of space-time at this dimension would be some kind of chaotic linear superposition of different topologies [112]—a picture in no way resembling a smooth manifold.

Whether or not it is meaningful to talk about the nature of space-time at such dimensions is not at all clear. But it if is *not* meaningful, then we *certainly* cannot refer to space-time as accurately constituting a smooth manifold. On the other hand, it may be argued that the smooth manifold picture is *adequate* for the discussion of all relevant physical processes. It is my personal view that this cannot *ultimately* be the case. I do not believe that a real understanding of the nature of elementary particles can ever be achieved without a simultaneous deeper understanding of the nature of space-time itself. But if we are concerned with a level of phenomena for which such an understanding is *not* necessary—and this will cover almost all of present-day physics—then the smooth manifold picture presents an (unreasonably!) excellent framework for the discussion of phenomena.

Let us now set aside the question of the submicroscopic structure of space-time and concentrate, instead, on its large-scale properties. In this case, we may imagine that the smooth manifold picture will be adequate and that its structure in the large can be obtained by piecing together smaller "locally Euclidean" patches, in the manner of the overlapping coordinate neighborhoods of differential geometry. Thus we might arrive at a topology for space-time, in the large, different from a Euclidean topology. Unfortunately, too little is known about the large-scale structure of the universe to enable us to make any statement with confidence concerning its global topology (apart, perhaps, from certain statements about its orientability). Thus, it *might* be that the topology of space-time on a large scale is not at all interesting.

Nevertheless, *irrespective* of this, it will be very worthwhile, here, to consider questions concerning topology of space-time manifolds in some generality. There are, in particular, two different but somewhat related special reasons for this. To appreciate, fully, the first of these, we must picture the universe in its four-dimensional entirety, rather than in terms of some three-dimensional spacelike section (the "present"). The one observational feature of the four-dimensional large-scale structure of the universe which *has* emerged with any semblance of certainty, is that at some "time" in the past (of the order of 10^{10} years ago) the material of the universe was apparently concentrated in a highly condensed and chaotic (hot) state. This follows from observed expansion of the universe and the equations of general relativity, if certain assumptions about the large-scale homogeneity of the universe are made. It also appears to follow more directly from the recent observation [25, 80a] of a general background of (electromagnetic) radiation permeating space. The present temperature of this radiation is about 3° absolute, which is what would be expected [33, 2] on the basis of a highly condensed general-relativistic "initial state." (The cooling down of the radiation to its present value would be an effect of the expansion of the universe.) Now, if we take completely seriously the smoothed-out cosmological models which are normally used in such calculations, we are actually presented with an initial *singular* state at which the curvature of space-time becomes infinite. In the neighborhood of the singularity, radii of curvature would become arbitrarily small—less than 10^{-13} cm; less, even, than 10^{-33} cm. Thus, we would expect the picture presented by our model to be inadequate at such levels, if only because of the reasons mentioned earlier for doubting the validity of the smooth manifold description of space-time at such dimensions.

But can we really trust our model in regions where the curvature even *remotely* approaches these values? We may expect that the deviations from homogeneity which are now present in the curvature of space-time (owing to irregularities in the matter density) could, if extrapolated backward to the early highly curved space-time regions, result in very great deviations from the smoothed-out picture. Indeed, is there reason to believe in the existence of a singularity at all? (We may take as a tentative definition of "singularity," a region at which curvatures have become so large, that the local physics becomes drastically altered—perhaps because of a breakdown in the smooth manifold picture of space-time.) Might it not be that when curvatures become only "moderately" large, the deviations from the *model* could be enormous, perhaps resulting even in a *topological* structure totally different from that of the model? It will be one of the main purposes of the later lectures in this series to give some precise results which strongly indicate (although, perhaps, they do not *quite* prove) the existence of singularities in space-time, in *generic* situations, on the basis of general relativity. In order to achieve these results it will be necessary to consider quite complicated possibilities of a somewhat topological nature—*even* if such complications are *not* actually realized in the universe!

This is the first of the two compelling reasons, alluded to above, for studying the subject of space-time topology. The second of these (leaving aside the question of relevance to the submicroscopic nature of space-time) concerns the question of gravitational collapse. The inherent *instability* of gravitation, as implied by general relativity, when too large concentrations of masses are present, is reflected in the existence of an initial singularity of the cosmological models. This instability becomes manifest again in the collapsing phase of those models for which the expansion does not continue idenfinitely and the universe returns to infinite curvature in a final singular state. But it is *not* necessary for the universe to be involved as a *whole*, for the effects of this instability to become manifest. In fact, even bodies which are not very much more massive than the sun may be expected to collapse catastrophically when they have exhausted their resources of internal energy. If the conditions are right, such a body will collapse past a "point of no return" at which (roughly speaking) the gravitational forces become so strong that even the *light* emitted by the body is dragged inward and no signal can escape to the outside world. Beyond this point the behavior of the body is much like the final phase of a collapsing universe (or, presumably, the time-reverse of the initial phase of an expanding one). Singularities in space-time would be expected to result, but (in the situation under consideration) these singularities would not be visible from the outside. Instead, there ensues something of the nature of a "hole" in space, into which objects can fall and out of which no object or signal can escape. Bearing in mind the mass range that might be likely to be involved, these "holes" could range in size from a few kilometers in diameter to a few times the dimension of the solar system. They would (apparently) be difficult—but not in principle impossible—to detect from outside. At the present time their probable existence is inferred theoretically rather than observationally, but in any case there are many intriguing questions of space-time topology involved in their study.

The nature of space-time is not quite what it "seems" to be—and in these respects it is not well understood at present. Undoubtedly, future theory will present us with many surprises and important new insights. But even the present theory of space-time (and by this I mean Einstein's 1916 theory of general relativity) contains surprises and insights which are only just beginning to be explored in any detail. There is surely a great wealth of results which may still be fairly readily obtained by use of the appropriate mathematical tools. I hope that by presenting a few of the results which have been so far obtained and by indicating some tools which have recently proved useful, I may be able to encourage experts in other fields to take a (possibly active) interest in this subject.

2 THE NATURE OF GENERAL RELATIVITY

The most important single lesson of relativity theory is, perhaps, that space and time are not concepts that can be considered independently of one another but must be combined together to give a *four*-dimensional picture of phenomena: the description in terms of *space-time*. Dynamics then becomes an aspect of geometry. It is somewhat instructive to see how this idea may be applied even in prerelativistic dynamical theories. Thus, we may compare the dynamics of Aristotle, Galileo, or Newton with special or general relativity, by expressing all five types of theory in space-time terms.

Let us consider five kinds of space-time, which I shall label as follows:

ARISTOTELIAN SPACE-TIME (2.1)
GALILEAN SPACE-TIME (2.2)
NEWTONIAN SPACE-TIME (2.3)
MINKOWSKIAN SPACE-TIME (2.4)
EINSTEINIAN SPACE-TIME (2.5)

In each case, the space-time will be a four-dimensional smooth manifold, but with some additional geometric structure which describes an important aspect of the dynamics. Any *point* of the space-time is really an "event," that is, something one pictures as a point in space, but with only an instantaneous existence. The history of a particle is a *curve*[1] in space-time, called the world line of the particle.

Aristotelian space-time is simply a product $E^3 \times E^1$, where E^n denotes the n-dimensional Euclidean space carrying the usual Euclidean metric and possessing the usual ($\frac{1}{2}n(n+1)$-dimensional) group of motions. The metric of E^3 defines spatial separation and that of E^1 defines time-difference. Thus, in accordance with Aristotelian dynamics, it is meaningful to speak of an absolute spatial separation between two events, even when they possess a nonzero time difference. In particular, the state of rest of a particle is to be distinguished from all other motions, by the fact that the spatial separation vanishes between any two points on the particle's world line. The seven-parameter transitive group of motions of Aristotelian space-time is the direct product of the Euclidean group on E^3 with that on E^1.

---

[1] Unless otherwise stated, a "curve" will not carry a parameter (cf. section 9).

Galilean and Newtonian space-times differ from Aristotelian space-time in that the spatial separation of two points is only well-defined if their time difference vanishes. On the other hand, the time difference is always well-defined. Thus, the structure is more accurately that of a *fiber bundle* over E^1 with fiber E^3, so that the "time" E^1 may be thought of as the *factor space* of the whole space by the fiber E^3. (The topology is, of course, still the same as in the Aristotelian case, but the bundle structure is different.[2]) The distinction between the Galilean and Newtonian space-times arises only after some further structure has been imposed. For this, we single out a special (six-parameter) family of curves in the space-time which we refer to as the "geodesics." (No extremal property of these curves is intended to be suggested here. We merely require that a unique such geodesic pass through each point of the space-time in each direction.) These geodesics (except for those lying in the E^3 fibers) will then be the world lines of particles in *inertial motion*. To obtain the geodesics for Galilean space-time, we merely postulate that they agree with the system of straight lines in some E^4, where the E^3 fibers are identified as a maximal system of parallel planes in E^4 and where the factor space E^1 is defined in the obvious way. The group of motions of the Galilean space-time which preserves this structure is the ten-parameter Galilei group.

The definition of the geodesics for Newtonian space-time is more complicated. The idea, here [20, 108, 109] is to treat Newtonian gravitational theory from the point of view presented by Einstein's general theory. Thus, we do *not* regard gravitation as a *true force*, but as an intertial or "fictional" force along with the acceleration, centrifugal, and Coriolis forces. This is possible[3] because of the Galilei–Einstein equivalence principle, which states the equality between inertial and passive gravitational mass. Then a particle moving under the action of gravity, but not influenced by any other force, is considered to be moving inertially and its world line is to be a

CHAPTER 36

---

[2] I am grateful to A. Trautman for clarifying this point for me.

[3] In fact, the principle of equivalence really also tells us that it is *necessary* to take this view of Newtonian gravitation. A constant gravitational field (in the Newtonian sense), throughout the whole of space, would be totally unobservable, all bodies being accelerated equally. The intrinsic physics would be the same as if the field did not exist. The situation is somewhat similar to that of electric potential. If we add (for example) a constant potential over the whole universe, the intrinsic physics is totally unaltered. Thus we think of Newtonian gravitational "field" as being really a type of potential. We have to *differentiate* to get the true field, namely, the curvature. We might imagine fixing the potentials canonically by requiring that they go to zero at infinity, but it is difficult to see how one would apply this in the actual universe, there being no evidence that the density of gravitating bodies falls off toward infinity.

geodesic. Many different Newtonian space-times are possible, corresponding to the various inequivalent gravitational fields. Galilean space-time is a special case, corresponding to zero gravitational field—or, equivalently, for example, to a field of *constant* gravitational acceleration defined throughout space-time. For, as in general relativity proper, the physically meaningful aspect of a gravitational field is contained only in the "tidal force" which results from a *nonuniform* gravitational acceleration field. These tidal forces find expression in a "curvature" for Newtonian space-time which describes the extent to which the intrinsic geodesic structure differs from that of the Galilean space-time (Fig. 1). The "field equations" of Newtonian space-time relate this curvature to the distribution of gravitating matter. These may be obtained by applying to limit $c \to \infty$ to the equations of general relativity (c being the velocity of light) but the details of this will not be entered into here.[4]

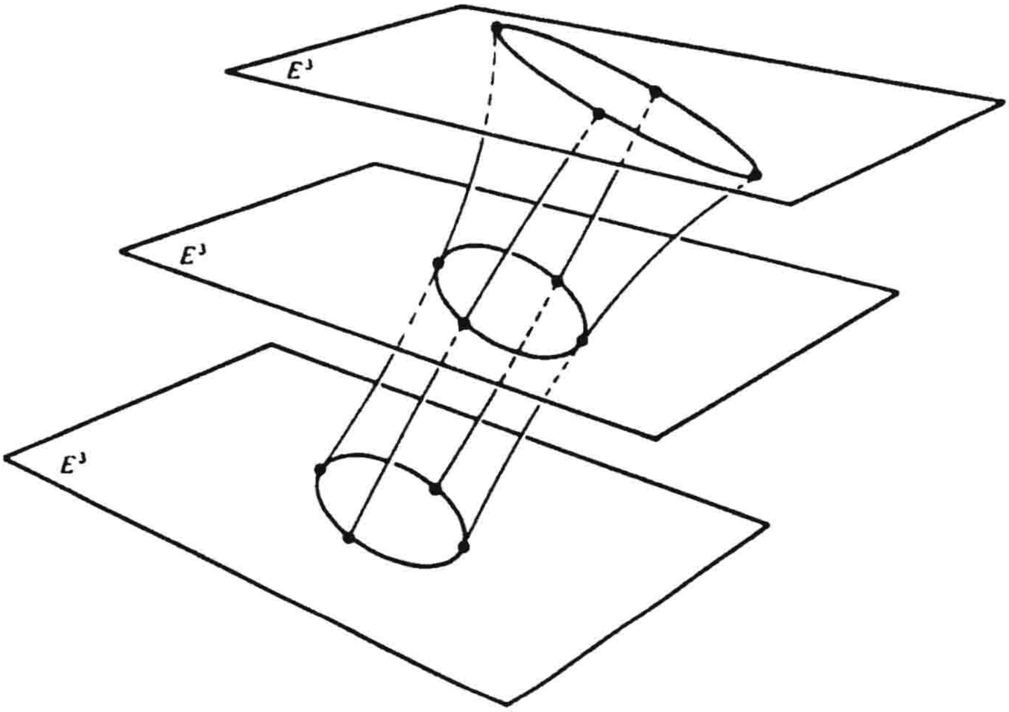


FIGURE 1. The curvature of Newtonian space-time is directly measured by geodesic deviation, that is, by tidal forces.

Minkowskian and Einsteinian space-times differ radically from the three previous cases in that *no* additive concept of time difference is defined between events. Instead, a pseudo-Riemannian[5] metric form ds^2, of hyperbolic normal signature (that is, $+, -, -, -$) is defined on the space-time. The time difference between two points A, B in the space-time *depends on the choice of world line connecting the points* and is given by the integral of ds along the world line:

---

[4] For the details of this, see Trautman [108, 109]. Trautman also considers the more general case where the Newtonian spaces of constant time need not be Euclidean E^3's. This enables "Newtonian cosmology" to be considered.

$$\tau = \int_A^B ds \tag{2.6}$$

For an allowable (that is, "timelike" or "null") world line, we have $ds^2 \geqslant 0$ everywhere along the curve, whence τ is, in fact, a real number. The quantity τ defines the time interval (*proper time*) between A and B as measured by an (idealized) clock whose world line is the given curve. Since proper time is now a path-dependent concept, we can return to the definition of a "geodesic" as an *extremal* path. The system of (timelike) geodesics thus obtained, defines (according to the theory) the inertial motions of particles. Unlike in the Galilean and Newtonian cases just considered, the inertial motions are now *fixed* once the behavior of (idealized) clocks has been specified.

The relation between Minkowskian and Einsteinian space-times mirrors that between Galilean and Newtonian space-times. Thus, Minkowskian space-time has a unique description (for example, its geodesic structure is the same as that of E^4, the set of timelike geodesics corresponding to geodesics in E^4 making an angle of less than 45° with some fixed direction) and it does not describe gravitation. It has a ten-parameter transitive group of motions, namely, the Poincaré (that is inhomogeneous Lorentz) group. On the other hand, many Einsteinian space-times exist, corresponding to the various inequivalent gravitational fields. As with the Newtonian space-time, it is the *tidal* gravitational force which has physical significance, this being present when there is a gradient of the "gravitational acceleration field." This tidal force is described in terms of deviations of the intrinsic geodesic structure from that of Minkowskian space-time, that is to say, by the *curvature* of the Einsteinian space-time [*cf.* (7.8)]. Einstein's field equations then describe how this space-time curvature is to be related to the density of matter (that is of stress–energy–momentum).

Before adopting the Einsteinian point of view, it is reasonable to question whether space-time actually possesses a uniquely and accurately defined pseudo-Riemannian structure. This depends on the existence of accurate clocks in nature, on the fact that such clocks behave locally according to the laws of special relativity and on the fact that for two coincident

CHAPTER 36

---

[5] The word "Riemannian" is generally reserved for the case of a positive-definite metric; thence the term "pseudo-Riemannian" for indefinite metrics. Sometimes the word "Lorentzian" is used for a pseudo-Riemannian manifold of signature $(+, -, \ldots, -)$.

clocks, the time rate they register is in a well-defined ratio which does not depend on their location in space-time nor on their histories. It seems that the existence of accurate clocks is deeply related to the quantum nature of matter. Ultimately this comes down to the fact that a natural frequency ν is associated (via Planck's constant h) with any mass m. For, combining Einstein's mass–energy law $E = mc^2$ with Planck's formula $E = h\nu$ (and choosing units, as we shall, henceforth, so that the velocity of light $c = 1$) we have

$$\nu = mh^{-1} \tag{2.7}$$

Any fundamental particle therefore defines a scale of time via its own rest mass m. The time-scale may be thought of as a series of "ticks" along the world line of the particle,[6] these being (by definition) a separation ν^{-1} apart. This gives us a definition of interval ds along the particle's world line and by allowing this world line to vary, we have ds for *any* timelike interval. Because of the very great local accuracy of special relativity, this then serves to define

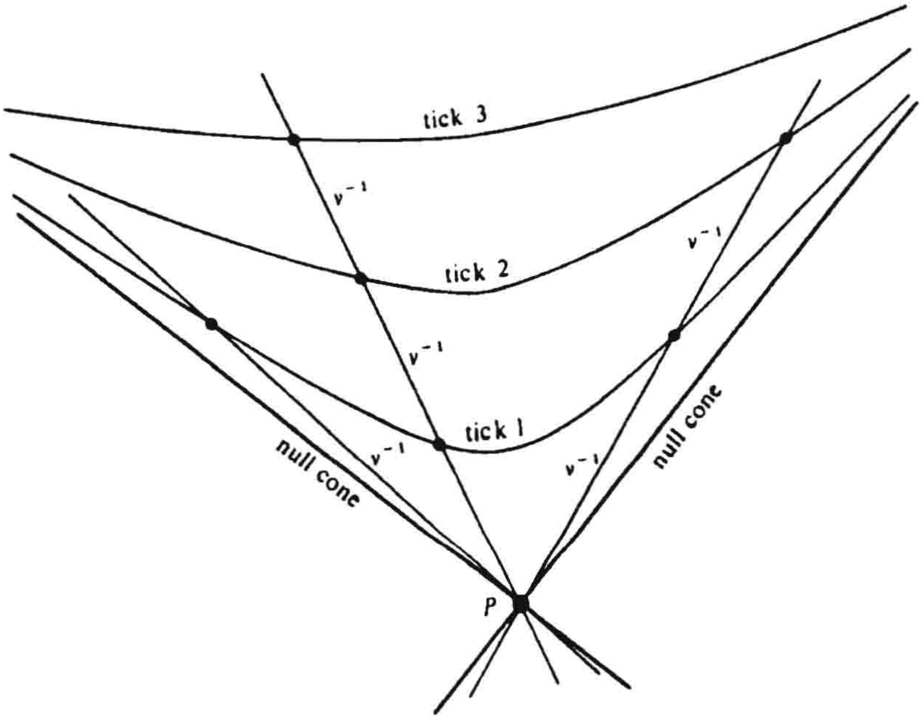


FIGURE 2. The ds defined by a particle moving in different timelike directions at a point P is consistent with a pseudo-Riemannian structure. This is because of the local accuracy of special relativity.

---

[6] In practice, the Planck frequencies of elementary particles are extremely high and cannot be used directly as clocks. The corresponding Planck frequencies for compound systems are even higher! Essentially the frequencies used in atomic or nuclear clocks are obtained from *differences* of masses (that is, "beats").

a pseudo-Riemannian structure for space-time (see Fig. 2). Finally, it appears that the pseudo-Riemannian structure we obtain is independent of the choice of particle that we use to define it. This is because the masses of elementary particles seem to be in a definite fixed ratio to one another, which is independent of their location in space-time or their histories. (If this were found not to be the case, then we should only have a uniquely defined *conformal* geometry, that is, a metric up to a local factor.)

So far, I have made no mention of spacelike separations, even though one normally thinks of Riemannian geometry in terms of distances rather than times. This is because a distance measurement is really a more complicated process than a time measurement. Given any two nearby points P, Q with a spacelike separation, we cannot just put a "ruler" between them to measure their separation. A ruler is not really an appropriate object for measuring distances between events. In space-time, a ruler appears as a two-dimensional timelike strip. Unless we arrange that the points P, Q on the two edges of the strip are "simultaneous in the ruler's rest frame," we shall get the wrong answer for the interval PQ (Fig. 3). In fact, the rigidity of a ruler is also a complicated matter and depends ultimately on the intersections between the "clocks" defined by its constituent atoms. To measure

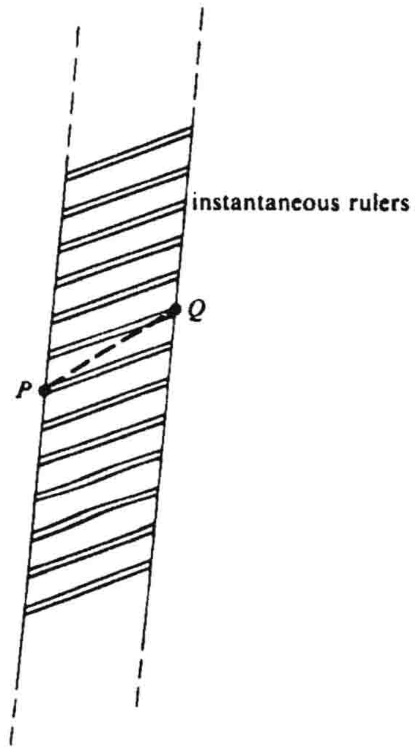


FIGURE 3. A ruler wrongly measuring a spacelike interval PQ. The events P, Q are not simultaneous in the rest-frame of the ruler.

spatial intervals, it is actually more direct to use a *clock*[7] and reflected light signals, such as in the manner indicated in Fig. 4. The spacelike interval PQ is then equal to the timelike interval between emission and reception of the signals (assuming the signals are emitted in opposite directions relative to the clock, that is, that the whole figure lies in a timelike plane).

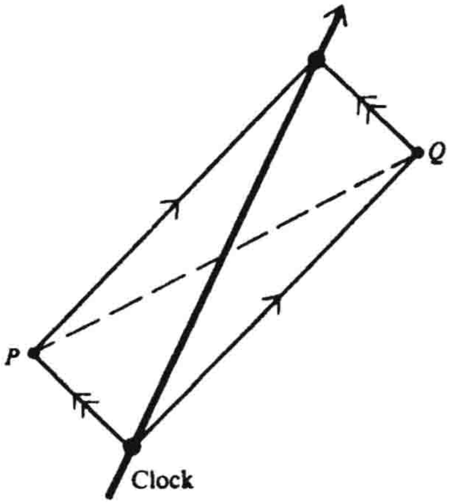


FIGURE 4. A clock correctly measuring the spacelike interval PQ by use of reflected light signals. (The whole figure must lie in a plane.)

Let us, then, agree that space-time, via accurate clocks, possesses a uniquely defined natural pseudo-Riemannian metric (of signature $+, -, -, -$). It is still reasonable to question whether the geodesics defined by this metric are likely to have anything to do with the inertial motions of particles. When general relativity was first put forward, the identity of the inertial motions with the timelike geodesics was assumed as a *postulate* of the theory. However, Einstein and Grommer [30] were later able to show,[8] that this property is actually a *consequence of Einstein's field equations* [*cf.* (7.1)] if certain reasonable assumptions are made about how a (test) particle is to be represented as a limitingly small distribution of energy–momentum. The argument is a delicate one, but it depends ultimately on the existence of a covariant conservation law for the energy tensor [*cf.* (8.1)]. Beyond this, it does not depend critically on the exact form of Einstein's field equations.

It is also possible to run the argument the other way, that is, to regard inertial motions as primary and to try and construct the metric as a secondary concept. According to an argument of Weyl [111] and of Marzke and Wheeler [56], if both the timelike geodesics (inertial particle world lines) and null geodesics (unscattered light rays) are known, then the metric of space-time can be constructed uniquely up to an overall factor. This, in a sense, has some advantage over the definition in terms of clocks, since the primary concepts do not depend on parts of physics (that is, quantum theory) which may be felt to be foreign to general relativity. My personal preference for regarding the clocks as more basic arises partly from the fact that one needs a number of apparently *ad hoc* assumptions on the geodesic structure of space-time in order that it be consistent with any metric at all, so it ceases to be clear that the structure of space-time is actually a pseudo-Riemannian one. In addition, it is my personal view that there *is* a deep connection between quantum theory and general relativity, so that it may actually be a mistake to attempt to build the subjects up separately.

---

[7] Synge [103] has particularly emphasized that the study of space-time geometry is essentially "chronometry."

[8] For an important later paper see [31]; for a more recent survey, see [36].

For general relativity to be a subject of relevance to nature, it is necessary that the pseudo-Riemannian structure of space-time be not only accurately defined, but also *nontrivial* (that is, with somewhere nonvanishing curvature). The fact that the space-time metric *is* curved[9] is actually a consequence of very primitive considerations [96] concerning the nonexistence of a *perpetuum mobile*. Consider a closed chain of buckets [9, p. 418] stretched between two pulleys, the axes of the two pulleys being fixed relative to the earth at different gravitational potentials (Fig. 5). Each bucket contains an atom which is capable of two states, a ground state and an excited state. The atoms on the left-hand side are all excited while those on the right are in the ground state. Now, the excited atoms, possessing more energy than the ground-state atoms, are also more massive. Therefore (by a very weak version of the principle of equivalence), they must also *weigh* more. We assume that there is no friction in the system, whence the chain begins to rotate in a counterclockwise direction. When each bucket reaches the bottom it is induced to give up its excess energy in the form of a photon which, by means of mirrors, is reflected back to the top atom. Now *if* the photon could be reabsorbed to excite the top atom we would have a *perpetuum mobile* since the process could be repeated indefinitely and energy could, in

CHAPTER 36

---

[9] The following argument will not show that space-time is *conformally* curved (that is, with a metric not expressible in the form $ds^2 = f(x, y, z, t)\ (dt^2 - dx^2 - dy^2 - dz^2)$). For this, the most direct argument is the bending of light by the sun. Space-time curvature seems to be about the only way of "bending" light without the different frequencies being bent differently (chromatic dispersion). In a conformally flat theory (for example, Nordström's [66] gravitational theory), there would be *no* resultant bending of light by gravitation (essentially because null geodesics are conformally invariant). It may be taken that the existence of *some* achromatic bending is experimentally established.

The most obvious manifestation of space-time curvature is the existence of planetary orbits. I do not choose to use this as the primary reason for believing in the curvature of space-time, because the connection of curved orbits with the behavior of clocks is not very direct, depending on the derivation of equations of motion from field equations on the metric (see above).

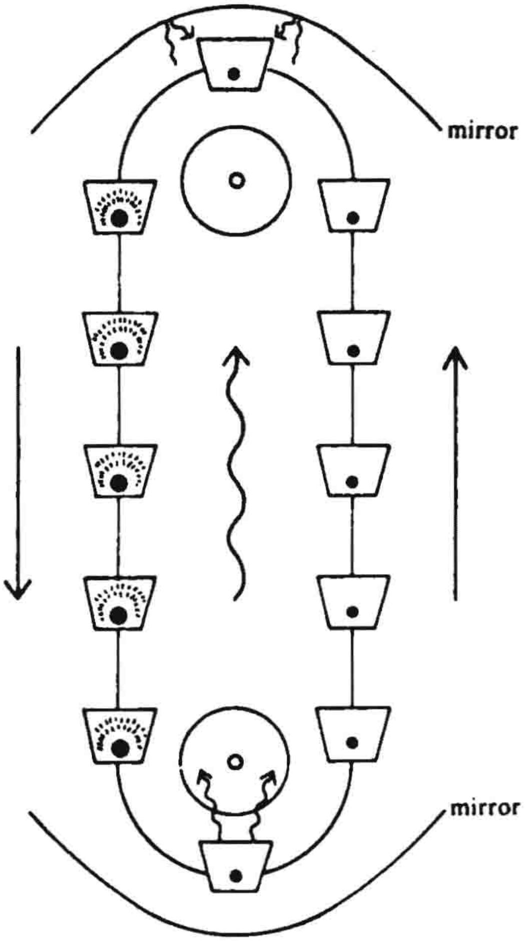


FIGURE 5. **Bondi's endless chain of buckets. The weight of the extra energy of the excited atoms would cause the chain to rotate.**

principle, be extracted from the continuing rotation of the chain. Thus, we must assume that the photon is *not* capable of being absorbed by the top atom. Presumably, if Newton had been presented with this situation, he would have had a simple answer: The photon is "weakened" by the time it reaches the top and possesses insufficient energy to excite the top atom. But *because of quantum theory*, we know that a photon cannot be *just* "weakened." By Planck's formula, it must also be of a *lower frequency* when it reaches the top than that required to raise the top atom to its excited state. Thus, in a well-defined sense, clocks must run more slowly at the bottom of the chain than at the top. Thus, the metric of space-time, as we have defined it, must differ, in a gravitational field, from the ordinary flat space-time form. (This is a highly idealized piece of apparatus, but it is of interest to note that this clock slowing effect has actually been directly observed on the earth, by Pound and Rebka [85] by use of the Mössbauer effect.) Note that the principle of equivalence is only invoked in an exceedingly weak form, namely, that energy just has *some* weight. (An extreme version of the experiment would be to employ π° mesons rather than atoms. These can decay entirely into photons, leaving the right-hand buckets empty. Thus, the argument will apply if π° mesons weigh anything at all!)

This argument is not, so far, quite sufficient to show that space-time is curved, however. The metric might still be flat but represented in an unusual way. From the spherical symmetry of the earth's field this is readily seen to be, in fact, impossible, since it would imply that the metric of the earth's field is related to a standard Minkowski space-time by transformations involving an acceleration outward from the center of the earth in all directions. The matter will not be pursued in further detail here, except that I feel it will be instructive just to examine the "uniform acceleration" transformation of Minkowski space-time explicitly. This illustrates quite clearly how a static situation *can* appear to involve a gravitational field, while the space-time is actually flat. This example also shares some features of the relation between the Schwarzschild and Kruskal representations of a true spherically symmetric gravitational field, as we shall see later ((10.1), (10.4); see also Bergmann [6a]).

Let x, y, z, t be standard Minkowskian coordinates, so that the metric takes the form

$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2 \tag{2.8}$$

Introduce new coordinates X, Y, Z, T related to the original ones in the range $z > |t|$; $Z > 0$, by

$$x = X,\ y = Y,\ z = Z \cosh T,\ t = Z \sinh T \tag{2.9}$$

(see Fig. 6). The metric now becomes

$$ds^2 = Z^2\, dT^2 - dX^2 - dY^2 - dZ^2 \tag{2.10}$$

The form (2.10) has the appearance of a static gravitational field with Z playing the role of a gravitational potential. The term $(Z\, dT)^2$ in (2.10) in place of the dt^2 seems to indicate that "time is running more slowly" near $Z = 0$. Any particle held "fixed" in the X, Y, Z, T system (that is with X, Y, Z constant) would experience a constant force in the direction of decreasing Z (since the particle's world line is "really" a constant acceleration "hyperbola" in the x, y, z, t system). A freely falling particle would appear, in the X, Y, Z, T system, to fall toward $Z = 0$ and then slow down and approach $Z = 0$ asymptotically (see broken curve in Fig. 6). There is an apparent "barrier" a $Z = 0$ which particles do not seem to be able to cross. That this is merely an effect of the inadequacy of the (X, Y, Z, T) coordinate patch becomes evident if we transform back to the form (2.8). It will be useful to bear this feature in mind when we come to consider the discussion of gravitational collapse given in Section 10.

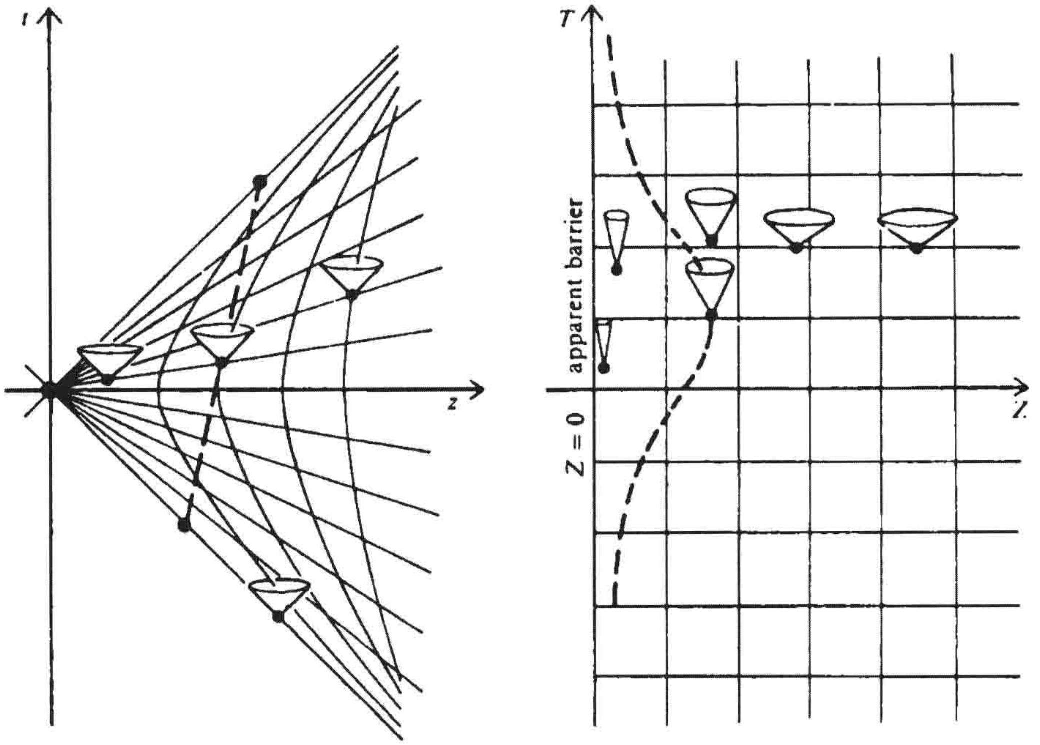


FIGURE 6. The uniform acceleration transformation. The apparent barrier at $Z = 0$ has no physical reality.

3 THE ABSTRACT INDEX NOTATION

When performing calculations in general relativity it is frequently necessary to operate with tensors of quite high valence.[10] Even such a basic quantity as the curvature tensor has valence four, and it possesses the familiar somewhat complicated symmetries. This makes it practically imperative that an index notation be employed for many calculations, so that the different connections between the quantities involved may be easily kept track of. It seems that there is a common feeling among mathematicians that such notations are to be avoided, presumably because of the connotation that their use entails explicit reference to a particular basis frame. However, when a physicist refers to "g_{ab}" or "$R^{a}{}_{bcd}$," I do not think that he usually means to be referring to a set of frame-dependent components but rather to a physical, frame-independent object which these components represent. But, the index notation allows a very convenient set of algebraic operations to be applied to such objects, which produce new objects—these operations being actually completely frame-independent. The algebraic operations are, in essence, extremely simple, but they also allow great flexibility in the building up of more complicated operations out of simple ones. It would seem a great pity to forbid oneself the use of such a powerful and flexible notation merely because of some uneasy feelings about summation conventions and dependence on special basis frames. What I shall present here is an entirely frame-independent algebra which allows one to calculate with indexed quantities exactly as before (but now with a clear conscience!) and which, by use of a notational device, even permits a greater freedom than before, when it comes to introducing coordinate systems and basis frames (compare Schouten [96a]). The advantages will be particularly apparent when we come to consider spinors in the next section.

---

[10] The term "valence" is used here in preference to "rank" because it is more descriptive and because the word "rank" has other connotations in the case of matrices.

Let us not be completely formal, so we may be able to save some time and complication. I hope the essential ideas will be clear. Consider a vector space $\mathbf{V}^\bullet$ over a field $\mathbf{F}$—or, more generally, we allow $\mathbf{V}^\bullet$ to be a module,[11] where $\mathbf{F}$ is a ring of suitable type (for example, the elements of $\mathbf{V}^\bullet$ could be vector fields[12] and those of $\mathbf{F}$, C^∞ functions, on a manifold). The idea will be to construct what is essentially the usual tensor product of $\mathbf{V}^\bullet$ a number of times with its dual[13] $\mathbf{V}_\bullet$ a number of times, but where, by use of indices, we can keep track of the effect of symmetries and contractions easily. This is done by simply mirroring the usual index notation (with summation conventions, etc.) but where now the indices $a, b, c, \ldots$ are *not* to be thought of merely as generic symbols standing, say, for $0, 1, 2, \ldots, N$, but as *abstract labels*. We shall require an infinite supply

$$a, b, c, \ldots, a_0, b_0, \ldots, a_1, b_1, \ldots, a_2, \ldots \tag{3.1}$$

of abstract labels, so that expressions of arbitrary length can be built up. Let $\mathbf{L}$ denote the set of labels (3.1). For any element ξ of $\mathbf{V}^\bullet$ and any label $x \in \mathbf{L}$, we shall allow ourselves to write a symbol ξ^x. As ξ ranges over the elements of $\mathbf{V}^\bullet$, the associated object ξ^x ranges over a corresponding set $\mathbf{V}^x$. It should be emphasized here that ξ^x is an entity in its own right and *not* the set of components of ξ in some frame. Now, since we wish to mirror the usual tensor rules for indexed quantities, we are not permitted to write $\xi^a + \eta^b$, but $\xi^a + \eta^a$ and $\xi^b + \eta^b$ will be both allowable. (We must think of ξ^a and ξ^b as *different* objects.) For any $\lambda \in \mathbf{F}$, we shall also be permitted to write $\lambda\xi^a$. Thus, each of $\mathbf{V}^a, \mathbf{V}^b, \ldots, \mathbf{V}^{a_0}, \ldots$ is a vector space or module canonically isomorphic with $\mathbf{V}^\bullet$.

---

[11] A module differs from a vector space in that the scalars form a ring with identity rather than a field. A ring differs from a field in that division by nonzero elements is not always possible.

[12] Here a "field" means a cross-section of the appropriate vector bundle over $\mathscr{M}$.

[13] The space of all linear mappings of the module $\mathbf{V}^\bullet$ into the ring $\mathbf{F}$.

It may be felt that it is unnatural to introduce an infinite number of isomorphic spaces, when actually we only have one space. But we may view the situation in a slightly different way. Each element of $\mathbf{L}$ is really just a kind of organizational marker which keeps a tag on a particular vector (etc) irrespective of where it may occur in an expression. Thus, ξ^x is just a *pair*, consisting of $\boldsymbol{\xi}$ together with the marker x. That is to say, it is an element of $\mathbf{V}^\bullet \times \mathbf{L}$. We then have $\mathbf{V}^a = \mathbf{V}^\bullet \times (a)$, $\mathbf{V}^b = \mathbf{V}^\bullet \times (b)$, etc. The vector space or module axioms will, of course, apply to each $\mathbf{V}^x$:

$$\begin{aligned}\xi^x + (\eta^x + \zeta^x) &= (\xi^x + \eta^x) + \zeta^x\\ \lambda(\xi^x + \eta^x) &= \lambda\xi^x + \lambda\eta^x\\ (\lambda + \mu)\xi^x &= \lambda\xi^x + \mu\xi^x\\ \lambda(\mu\xi^x) &= (\lambda\mu)\xi^x\\ 1\xi^x &= \xi^x\\ 0\xi^x &= 0\eta^x\end{aligned} \tag{3.2}$$

Here λ, μ, $1, 0 \in \mathbf{F}$ with 1 and 0 being the multiplicative and additive identities, respectively. We also have $\xi^x + \eta^x = \eta^x + \xi^x$ (expanding $(1 + 1)(\xi^x + \eta^x)$) and $\xi^x + (-\xi^x) = 0$ (writing $-\xi^x$ for $(-1)\xi^x$ and 0 for $0\eta^x$).

The *dual* space $\mathbf{V}_\bullet$ will also have an infinite number of canonically isomorphic copies: $\mathbf{V}_a, \mathbf{V}_b, \ldots, \mathbf{V}_{a_0}, \ldots$. We may think of $\mathbf{V}_x$ as being the dual space of $\mathbf{V}^x$ for each $x \in \mathbf{L}$. The elements of $\mathbf{V}_x$ are linear mappings of $\mathbf{V}^x$ into $\mathbf{F}$. Thus, for $\theta_x \in \mathbf{V}_x$ we have

$$\theta_x(\xi^x + \eta^x) = \theta_x\xi^x + \theta_x\eta^x \tag{3.3}$$

$$\theta_x(\lambda\xi^x) = \lambda(\theta_x\xi^x) \tag{3.4}$$

where the effect of the mapping θ_x on ξ^x is written simply $0_x\xi^x$. We shall also allow this to be written in the reverse order: $0_x\xi^x = \xi^x 0_x$. We require

$$0_a\xi^a = 0_b\xi^b = \cdots = \theta_x\xi^x = \cdots \tag{3.5}$$

Now each of $\mathbf{V}_a$, $\mathbf{V}_b$, ... will be vector space or module where $\lambda 0_x$ and $\theta_x + \phi_x$ are defined by

$$(\lambda\theta_x)\xi^x = \lambda(\theta_x\xi^x) \qquad (\theta_x + \phi_x)\xi^x = 0_x\xi^x + \phi_x\xi^x \tag{3.6}$$

The idea will be to use the elements of $\mathbf{F}$, $\mathbf{V}^a$, $\mathbf{V}^b, \ldots, \mathbf{V}_a, \mathbf{V}_b, \ldots$ to generate our algebra. To see what the rules of this algebra should be, we must recall what the rules are for the ordinary tensor index notation. We note, for instance, that products such as $\xi^a\eta^b$ will be permitted, whereas $\xi^a\eta^a$ will not. Furthermore, the allowable products must be *commutative*:

$$\xi^a\eta^b = \eta^b\xi^a \tag{3.7}$$

but in general $\xi^a\eta^b \neq \eta^a\xi^b$. The requirement (3.7) shows us that while *in essence* $\xi^a\eta^b$ is just the tensor product of elements $\xi^a \otimes \eta^b$, we cannot simply *identify* $\xi^a\eta^b$ with $\xi^a \otimes \eta^b$. For tensor products, according to the strict technical definition, are not commutative. Here we are allowed to define a commutative version of a tensor product, essentially for the reason that $\xi^a\eta^a$ is not defined. In a product $\xi^a\eta^b$, it is the *labels a* and *b* that tell us which factor is which, not the ordering of the factors. One method (suggested to me by S. Mac Lane) of precisely defining the type of product used here is to take the *symmetric algebra* [54a] on the direct sum $\mathbf{V}^a \oplus \mathbf{V}_a \oplus \mathbf{V}^b \oplus \mathbf{V}_b \oplus \ldots$ and then, for each pair of disjoint (finite) sets of elements of $\mathbf{L}$, say a, p, r and b, m, we select the corresponding subspace $\mathbf{V}^{apr}_{bm}$ spanned by the elements of the form

$$\xi^a\eta^p\zeta^r\theta_b\phi_m \tag{3.8}$$

The general element of $\mathbf{V}^{abr}_{bm}$ will be a linear combination of expressions like (3.8). The construction so made, ensures that each product (3.8) is fully commutative and also that the various distributive laws hold (for example, $\psi^{apr}_m(\theta_b + \chi_b) = \psi^{apr}_m\theta_b + \psi^{apr}_m\chi_b$ with $\psi^{apr}_m \in \mathbf{V}^{apr}_m$). There is no significance in the ordering of a, p, r in $\mathbf{V}^{apr}_{bm}$ or of b, m. Thus $\mathbf{V}^{apr}_{bm} = \mathbf{V}^{par}_{bm} = \mathbf{V}^{arp}_{mb}$, etc. However, the ordering of the indices for an element ρ^{apr}_{bm} *is* significant.

Every element $\rho^{apr}_{bm} \in \mathbf{V}^{apr}_{bm}$ is a linear combination of commutative products of the type (3.8):

$$\rho^{apr}_{bm} = \sum_{i=1}^{M} \overset{(i)}{\lambda}\, \overset{(i)}{\xi^a}\, \overset{(i)}{\eta^p}\, \overset{(i)}{\zeta^r}\, \overset{(i)}{\theta_b}\, \overset{(i)}{\phi_m} \tag{3.9}$$

but there will be many ways of expressing ρ^{apr}_{bm} as such. A convenient criterion for the equality of two expressions (3.9) for the modules that we are interested in is that for *every* choice $\alpha_a \in \mathbf{V}_a$, $\beta_p \in \mathbf{V}_p$, $\gamma_r \in \mathbf{V}_r$, $\sigma^b \in \mathbf{V}^b$, $\tau^m \in \mathbf{V}^m$, the scalar

$$\rho^{apr}_{bm}\alpha_a\beta_p\gamma_r\sigma^b\tau^m = \sum_{i=1}^{M} \overset{(i)}{\lambda}\,(\overset{(i)}{\xi^a}\alpha_a)(\overset{(i)}{\eta^p}\beta_p)(\overset{(i)}{\zeta^r}\gamma_r)(\overset{(i)}{\theta_b}\sigma^b)(\overset{(i)}{\phi_m}\tau^m) \tag{3.10}$$

should be the same for both expressions. From this all the algebraic properties will follow. (Note, that in general $\rho^{arp}_{bm} \neq \rho^{par}_{bm} \neq \rho^{apr}_{mb}$, etc.)

The entire tensor system $\{\mathbf{V}\}$ consists of all of the $\mathbf{V}^{x\ldots z}_{u\ldots w}$, including $\mathbf{V} = \mathbf{F}$:

$$\{\mathbf{V}\} = (\mathbf{V}, \mathbf{V}^a, \mathbf{V}^b, \ldots, \mathbf{V}_a, \mathbf{V}_b, \ldots, \mathbf{V}^{ab}, \ldots, \mathbf{V}^{x\ldots z}_{u\ldots w}, \ldots)$$

There are four basic operations on $\{\mathbf{V}\}$, namely

ADDITION: $\mathbf{V}^{x\ldots z}_{u\ldots w} \times \mathbf{V}^{x\ldots z}_{u\ldots w} \rightarrow \mathbf{V}^{x\ldots z}_{u\ldots w}$ (3.11)

MULTIPLICATION: $\mathbf{V}^{a\ldots d}_{p\ldots r} \times \mathbf{V}^{x\ldots z}_{u\ldots w} \rightarrow \mathbf{V}^{a\ldots dx\ldots z}_{p\ldots ru\ldots w}$ (3.12)

INDEX SUBSTITUTION: $\mathbf{V}^{x\ldots z}_{u\ldots w} \rightarrow \mathbf{V}^{f\ldots h}_{k\ldots m}$ (3.13)

(a, b)-CONTRACTION: $\mathbf{V}^{ax\ldots z}_{bu\ldots w} \rightarrow \mathbf{V}^{x\ldots z}_{u\ldots w}$ (3.14)

In (3.11), (3.12), and (3.14), the differently denoted index letters appearing are all assumed to be different elements of **L**. In (3.13), the elements $x, \ldots, z, u, \ldots, w$ of **L** are all distinct and so are $f, \ldots, h, k, \ldots, m$. Otherwise they are unrestricted except that $x, \ldots, z$ and $f, \ldots, h$ are equal in number and that $u, \ldots, w$ and $k, \ldots, m$ are equal in number. Addition and multiplication are defined the obvious way. Index substitution is induced simply by a permutation applied to **L**. (The validity of any equation is unaffected by a permutation of the elements of **L**.) To define contraction, we consider, for example, the (p, b)-contraction: $\mathbf{V}^{par}_{bm} \rightarrow \mathbf{V}^{ar}_{m}$ as applied to the element $\rho^{apr}_{bm} \in \mathbf{V}^{par}_{bm}$ given by (3.9). The result is

$$\rho^{axr}_{xm} = \sum_{\mathrm{i}=1}^{M} \overset{(\mathrm{i})}{\lambda} (\overset{(\mathrm{i})}{\eta}{}^{x} \overset{(\mathrm{i})}{\theta}{}_{x}) \overset{(\mathrm{i})}{\xi}{}^{a} \overset{(\mathrm{i})}{\zeta}{}^{r} \overset{(\mathrm{i})}{\phi}{}_{m} \tag{3.15}$$

We have $\rho^{axr}_{xm} \in \mathbf{V}^{ar}_{m}$, so the x labels are "dummies" which do not contribute to the total valence type.

It may be verified algebraically that all the usual tensor rules[14] follow from the above constructions. Thus, addition gives an Abelian group structure to each $\mathbf{V}^{x \cdots z}_{u \cdots w}$. Multiplication is commutative and distributive over addition. Contraction appropriately commutes with addition, with multiplication, and with other contractions. The contraction of a zero element is again a zero element. (If we use equality between the scalars (3.10) as the definition of equality between formal expressions (3.9), then this last property follows from the property that any matrix over **F** whose square vanishes also has vanishing trace. This property holds in the cases which interest us here, but would not be true for certain rings of finite characteristic.)

So far, the question of a basis frame for $\mathbf{V}^{\bullet}$ has not even arisen. However, it is often convenient to work with basis frames and we shall need a notation to be able to distinguish basis indices from the abstract labels. I shall adopt the convention that *German* indices $\mathfrak{a}, \mathfrak{b}, \ldots, \mathfrak{a}_0, \ldots$ will denote the numbering of basis elements in the standard way, that is, each of $\mathfrak{a}, \mathfrak{b}, \ldots$ denotes one of the integers $0, 1, \ldots, N$ [for an $(N + 1)$-dimensional space]. The use of German indices will be to remind us of two things; first that a choice of a (possibly arbitrary) basis frame is involved in any expression containing such indices, with a consequent loss of covariance; and secondly that the Einstein *summation convention* is being used whenever repeated indices occur in a term in an expression. Now, let $\boldsymbol{\delta}_0, \boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_N \in \mathbf{V}^{\bullet}$ be a basis for $\mathbf{V}^{\bullet}$ (assuming finite-dimensionality) and let $\boldsymbol{\delta}^0, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N \in \mathbf{V}_{\bullet}$ be the corresponding dual basis. For $x \in \mathbf{L}$ we have canonical images in $\mathbf{V}^x$ and $\mathbf{V}_x$:

$$\delta^x_0 \ldots, \delta^x_N \in \mathbf{V}^x, \qquad \delta^0_x, \ldots, \delta^N_x \in \mathbf{V}_x \tag{3.16}$$

---

[14] See any standard work on classical tensor calculus, for example, [100].

We may use the generic symbols $\delta^x_{\mathfrak{x}} \in \mathbf{V}^x$, $\delta^{\mathfrak{x}}_x \in \mathbf{V}_x$, so the basis orthogonality relation takes the form

$$\delta^x_{\mathfrak{y}}\delta^{\mathfrak{x}}_x = \delta^{\mathfrak{x}}_{\mathfrak{y}} \tag{3.17}$$

where $\delta^{\mathfrak{x}}_{\mathfrak{y}}$ is the ordinary Kronecker delta symbol. (No relation between x and $\mathfrak{x}$ is to be implied by the notation.) We can also define an element δ^x_y of $\mathbf{V}^x_y$ by

$$\delta^{\mathfrak{x}}_y\delta^x_{\mathfrak{x}} = \delta^x_y \tag{3.18}$$

The quantity δ^x_y satisfies the usual properties

$$\xi^{\cdots y \cdots}_{\cdots\ \cdots}\delta^x_y = \xi^{\cdots x \cdots}_{\cdots\ \cdots} \qquad \eta^{\cdots\ \cdots}_{\cdots x \cdots}\delta^x_y = \eta^{\cdots\ \cdots}_{\cdots y \cdots} \tag{3.19}$$

and is clearly actually independent of the choice of basis, by virtue of this. The *components*, with respect to the basis, of any element $\zeta^{x \cdots z}_{u \cdots w} \in \mathbf{V}^{x \cdots z}_{u \cdots w}$ are given by

$$\zeta^{\mathfrak{x} \cdots \mathfrak{z}}_{\mathfrak{u} \cdots \mathfrak{w}} = \zeta^{x \cdots z}_{u \cdots w}\delta^u_{\mathfrak{u}} \ldots \delta^w_{\mathfrak{w}}\delta^{\mathfrak{x}}_x \ldots \delta^{\mathfrak{z}}_z \tag{3.20}$$

Conversely, to express $\zeta^{x \cdots z}_{u \cdots w}$ in terms of its components, we simply write

$$\zeta^{x \cdots z}_{\mathfrak{u} \cdots \mathfrak{w}} = \zeta^{\mathfrak{x} \cdots \mathfrak{z}}_{\mathfrak{u} \cdots \mathfrak{w}}\delta^{\mathfrak{u}}_u \ldots \delta^{\mathfrak{w}}_w\delta^x_{\mathfrak{x}} \ \ldots \delta^z_{\mathfrak{z}} \tag{3.21}$$

One aspect of this notation which is an advantage in certain contexts is that we may convert some indices into component form and leave others as abstract labels:

$$\rho^{a\mathfrak{p}r}_{\mathfrak{b}m} = \rho^{apr}_{bm}\delta^{\mathfrak{p}}_p\delta^b_{\mathfrak{b}} \in \mathbf{V}^{ar}_m \tag{3.22}$$

Generally, all *algebraic* relations will be unaffected by whether or not an index is in German type. But when we consider covariant derivatives in the next section, we shall see that an important *formal* difference arises between the ways the two types of index are treated—in addition to the present merely conceptual difference.

An elementary, but important, property of the type of algebraic structure that we have built up here, is that we can sometimes embed one such structure in another by the device of *grouping together indices*. Thus, we may consider a new labeling set $\hat{\mathbf{L}}$, say, whose elements are (disjoint) subsets of elements of $\mathbf{L}$. For example, we could put $\alpha = abc$, $\beta = def$, $\gamma = ghi$, etc., where $\mathbf{L}$ is divided exhaustively into disjoint triplets, these triplets being the elements of $\hat{\mathbf{L}}$. It is clear that, in this case, the sets $\mathbf{V}^\alpha = \mathbf{V}^{abc}$, $\mathbf{V}^\beta = \mathbf{V}^{def}, \ldots, \mathbf{V}_\alpha = \mathbf{V}_{abc}, \ldots, \mathbf{V}^{\alpha\beta}_\gamma = \mathbf{V}^{abcdef}_{ghi}, \ldots$ will satisfy just the same rules as before. More complicated groupings are also possible.

We may also consider systems where the labeling set contains elements of different "types." The only modifications of our scheme would then come in index substitution (where only labels of the *same* type may be substituted for one another) and in contraction, where only an upper and lower indices of the *same* type may be contracted together. When we consider spinors in the next section we shall see an example of this kind of system. We shall have labels of two different types (related to each other via an operation of complex conjugation).

4 SPACE-TIMES WITH SPINOR STRUCTURE

In order to discuss the structure of space-time in detail, we shall need a definition of a differentiable manifold. I shall use the definition as given by Chevalley, *cf.* Nomizu [65], which gives the differentiable structure of a manifold $\mathscr{M}$ entirely in terms of the C^∞ real-valued functions on $\mathscr{M}$. I have chosen to use this definition here, not because of any special intrinsic mathematical merit, but because I feel it reflects most nearly, the role that is played by coordinate systems in physics. Thus, a coordinate system in space-time is actually a set of four real-valued functions on the space-time such that, in some neighborhood, knowledge of the values of these functions will fix the point in a "smooth" manner.

Let us start with space-time $\mathscr{M}$ as a topological space. A set of subsets of $\mathscr{M}$, referred to as the *open* sets, is singled out subject to:

THE INTERSECTION OF ANY TWO OPEN SETS IS OPEN. (4.1)

THE UNION OF ANY SET OF OPEN SETS IS OPEN. (4.2)

THE HAUSDORFF CONDITION HOLDS (THAT IS, FOR ANY TWO POINTS OF $\mathscr{M}$, TWO DISJOINT OPEN SETS EXIST EACH CONTAINING ONE OF THE POINTS). (4.3)

$\mathscr{M}$ IS CONNECTED (THAT IS, NOT THE UNION OF TWO DISJOINT NONEMPTY OPEN SETS). (4.4)

THERE IS A COUNTABLE NUMBER OF OPEN SETS, OF WHICH EVERY OPEN SET ON $\mathscr{M}$ IS A UNION. (4.5)

The concept of an open set is a very nonphysical one, but it is essential to the mathematical idea of a manifold. To dispense with the open set concept we would have to change the theory. The question of defining open sets in terms of the more physical notion of *causality* will be touched upon in Section 11.

We now single out a certain set T, of real-valued functions on $\mathscr{M}$, which will turn out to be the C^∞ real functions on the space-time $\mathscr{M}$. We can characterize the differentiable structure of $\mathscr{M}$ by the following axioms on T:

IF $f_1, f_2, \ldots, f_n \in \mathrm{T}$ AND IF ϕ IS ANY REAL-VALUED C^∞-FUNCTION OF n VARIABLES (IN THE ORDINARY SENSE), THEN $\phi(f_1, f_2, \ldots, f_n) \in \mathrm{T}$; (4.6)

IF f IS SUCH THAT, FOR EACH $P \in \mathscr{M}$ THERE EXISTS AN ELEMENT $f_{(P)} \in \mathrm{T}$, WHICH AGREES WITH f THROUGHOUT AN OPEN SET CONTAINING P, THEN $f \in \mathrm{T}$; (4.7)

FOR EACH $P \in \mathscr{M}$, THERE IS AN OPEN SET $\mathscr{S}$ CONTAINING P AND FOUR ELEMENTS $x^0, x^1, x^2, x^3 \in \mathrm{T}$ SUCH THAT THE MAP WHICH CARRIES $Q \in \mathscr{S}$ INTO $(x^0(Q), x^1(Q), x^2(Q), x^3(Q))$ IS A HOMEOMORPHISIM ONTO AN OPEN SUBSET OF $\mathbf{R}^4$, AND SUCH THAT ANY $f \in \mathrm{T}$ AGREES, ON $\mathscr{S}$, WITH SOME C^∞-FUNCTION OF $x^0, \ldots, x^3$. (4.8)

The property (4.8) expresses what is required of the functions $x^0, \ldots, x^3$ on $\mathscr{M}$ that they should be adequate as *local coordinates* about P.

We can define a C^∞ contravariant vector field ξ on $\mathscr{M}$ (that is, a C^∞ cross section of the tangent bundle of $\mathscr{M}$) as a *derivation* on the algebra T (over the constants) that is, a mapping $\xi\colon \mathrm{T} \to \mathrm{T}$ such that

$$\xi(f+g) = \xi(f) + \xi(g) \tag{4.9}$$

$$\xi(fg) = f\xi(g) + g\xi(f) \tag{4.10}$$

$$\xi(k) = 0 \text{ if } k \text{ is constant on } \mathscr{M} \tag{4.11}$$

The set $\mathrm{T}^\bullet$ of such ξ's is clearly a module over T, with $(\xi + \eta)(f) = \xi(f) + \eta(f)$, $(\lambda\xi)(f) = \lambda\xi(f)$ ($\lambda \in \mathrm{T}$; $\xi, \eta \in \mathrm{T}^\bullet$). It can be shown that, in terms of local coordinates $x^0, \ldots, x^3$ in an open set $\mathscr{S} \in \mathscr{M}$, any $\xi \in \mathrm{T}^\bullet$ has the form

$$\xi \equiv \xi^a \frac{\partial}{\partial x^a} \tag{4.12}$$

throughout $\mathscr{S}$. Thus, we can think of ξ^a as the components of ξ in the coordinate system x^a on $\mathscr{S}$. (When we pass to another local coordinate patch, the components ξ^a transform in the overlap region in the familiar way.)

At this point we could introduce, via $\mathbf{V}^\bullet = \mathrm{T}^\bullet$, $\mathbf{F} = \mathrm{T}$, the system of tensors with labels, as in Section 3, and we should arrive at a coordinate-free "abstract index" version of the familiar tensor calculus. What I propose to do here is something slightly different. I wish to present a somewhat unusual development of space-time theory. The reason for this, touches on what is perhaps an essential difference between mathematics and physics. In mathematics, one is interested in formalisms of maximum generality, so that any results obtained will have the widest possible application. Sometimes these applications are very far removed from the original topic, and great conceptual unifications may be thereby achieved. In physics, the aim is somewhat different. We are given a very specific (?) structure which we do not adequately understand—namely, the universe. Certain aspects of this structure are mirrored very accurately (in a miraculous way!) by certain mathematical models. However, we know, for various reasons, that these models are not quite (?) right. Thus, we are always interested in altering the mathematical theory. Now a formalism for physics which contains even an arbitrary parameter (such as the *dimension* of differentiable manifold, for example) is in a sense *too* general. We search for a formalism which is *specific* rather than general. Sometimes, by reformulating an old theory in an unusual (though mathematically equivalent) way, previously unexpected possibilities for modifying the theory may appear as mathematically natural. An example of this would be Newtonian gravitational theory. In the normal version of the theory, one might, for example, envisage changing the exponent -2 in the inverse square law to some other power. (In fact, such alterations were tried, in order to explain the motion of Mercury, but this does not work!) On the other hand, if Newtonian theory had been reformulated as a *space-time theory* in the way indicated in Section 2, then a quite different alteration—namely, to some form of general relativity—would have appeared as mathematically natural.

What I propose to do here is introduce a formalism for the description of space-time which makes its *spinor* structure[15] emerge as more basic even than its pseudo-Riemannian structure. In fact, the particular dimensionality and signature $(+, -, -, -)$, will have to be built in at the outset. Thus, if a modification of the present-day differentiable manifold picture of space-time were to emerge, which depended *essentially* on the existence of such a spinor structure, then the dimension and signature of our space-time would be one of the consequences of the theory.

The formalism will be based on the isomorphism[16] between the group $SL(2, \mathbb{C})$ of unimodular (2×2) complex matrices, and the twofold (universal) covering of the connected component of the Lorentz group $O(1, 3)$. To express this isomorphism in its most direct form, consider the general (2×2) Hermitian matrix, which I shall write as

$$\begin{pmatrix} u^{00'} & u^{01'} \\ u^{10'} & u^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} u^0 + u^1 & u^2 + iu^3 \\ u^2 - iu^3 & u^0 - u^1 \end{pmatrix} \tag{4.13}$$

If the matrix (4.13) is multiplied on the left by a unimodular (2×2) complex matrix $(t^{\mathfrak{A}}{}_{\mathfrak{B}})$ and on the right by the conjugate transpose of this matrix, then both the Hermiticity and the value of the determinant of (4.13) will be preserved. Thus, we obtain a linear transformation of (u^0, u^1, u^2, u^3) which preserves both the reality of u^a and the form

$$g_{ab} u^a u^b = (u^0)^2 - (u^1)^2 - (u^2)^2 - (u^3)^2 \tag{4.14}$$

---

[15] I am using the phrase "spinor structure" rather than "spin structure," to emphasize that in addition to being a spin manifold, $\mathscr{M}$ must possess a structure defined by the *particular* type of spinor system used here.

[16] The $(2-1)$ isomorphism between $SU(2, 2)$ and the connected component of $O(2, 4)$ can also be used in the study of space-time structure (*cf.* [79, 80]).

where

$$(g_{\mathfrak{ab}}) = \mathrm{diag}\,(1,\ -1,\ -1,\ -1) \tag{4.15}$$

This is therefore a Lorentz transformation on $u^{\mathfrak{a}}$:

$$u^{\mathfrak{a}} \rightarrow L^{\mathfrak{a}}_{\ \mathfrak{b}}\, u^{\mathfrak{b}} \tag{4.16}$$

which is continuous with the identity (since $t^{\mathfrak{A}}{}_{\mathfrak{B}}$ is continuous with $\delta^{\mathfrak{A}}_{\mathfrak{B}}$). Conversely, it follows (for a variety of reasons) that any such transformation (4.16) arises from precisely *two* unimodular matrices $(t^{\mathfrak{A}}{}_{\mathfrak{B}})$, namely, $(t^{\mathfrak{A}}{}_{\mathfrak{B}})$ and $(-\,t^{\mathfrak{A}}{}_{\mathfrak{B}})$. Thus, (4.16) is equivalent to

$$u^{\mathfrak{A}\mathfrak{A}'} \rightarrow t^{\mathfrak{A}}{}_{\mathfrak{B}}\, \bar{t}^{\mathfrak{A}'}{}_{\mathfrak{B}'}\, u^{\mathfrak{B}\mathfrak{B}'} \tag{4.17}$$

(The notation, here, is to regard $\mathfrak{A}$ and $\mathfrak{A}'$ as *distinct* letters, so, in particular, no summation is implied between them. Thus, (4.17) is *four* relations. When complex conjugates are taken, unprimed indices become primed and primed indices become unprimed. Thus, $\bar{t}^{0'}{}_{1'}$ is the complex conjugate of $t^{0}{}_{1}$, for example. Capital German indices run over the two values 0, 1 or 0′, 1′. Lower case German indices will run over 0, 1, 2, 3.)

The usual way to think of the components of a world vector in a pseudo-orthonormal frame is as a linearly ordered array $(u^0,\ u^1,\ u^2,\ u^3)$. But alternatively, the matrix version (4.13) would be just as good, and the idea here will be to regard this matrix description as *more* basic. In this way, a world vector appears as *not* the most elementary "vectorial" quantity in space-time, since it is now a "divalent" object and we would expect that there might also be "univalent" (two-component) objects with some space-time meaning. In fact, this turns out to be the case. These "univalent" objects will be called *spin-vectors*. The idea, then, will be to take as our basic module (the "$\mathbf{V}^{\bullet}$" of Section 2) not the set of world-vector fields $\mathbf{T}^{\bullet}$, but the C^{∞} *spin-vector* fields[17] $\mathbf{S}^{\bullet}$ on $\mathcal{M}$. The fact that I have not yet defined a spin-vector in space-time terms will not matter. The procedure is more the other way around. The algebra generated by the spin-vectors will serve, instead, to define the structure of the space-time. Once we have this structure we shall then be able to go back and interpret the spin-vectors in more familiar terms. We shall, in fact, end up with a remarkably complete interpretation of a spin vector in "physical" terms (*cf.* Section 5).

Now what intrinsic structure should our module $\mathbf{S}^{\bullet}$ of spin-vector fields possess? At each point $P \in \mathcal{M}$, we require the fiber $\mathbf{S}^{\bullet}(P)$ above P to be a two-dimensional complex vector space ("spin space"). The symmetry of this space is to be given by the group $SL(2,\ \mathbf{C})$, so that the local Lorentz group symmetry of space-time will be reflected in spin space symmetry. If ω^0, ω^1 are the components of $\omega \in \mathbf{S}^\bullet(P)$ in some basis frame, then an (active) symmetry transformation of $\mathbf{S}^\bullet(P)$ will be given by

---

[17] The system $\mathbf{S}^{\bullet}$ will turn out to be the C^{∞} cross sections of the bundle of spin-vectors over $\mathcal{M}$, regarded as a module over the C^{∞} complex valued functions on $\mathcal{M}$.

$$\omega^{\mathfrak{A}} \rightarrow t^{\mathfrak{A}}{}_{\mathfrak{B}}\, \omega^{\mathfrak{B}} \tag{4.18}$$

(keeping the basis frame fixed), the matrix $(t^{\mathfrak{A}}{}_{\mathfrak{B}})$ being complex and unimodular. Comparing (4.18) with (4.17) and (4.16) [via (4.13)] we see that if we regard the world-vector space $\mathbf{T}^\bullet(P)$ as being actually the tensor product of $\mathbf{S}^\bullet(P)$ with its complex conjugate, then the allowable (active) symmetries of $\mathbf{T}^\bullet(P)$ become just Lorentz transformations which do not involve space reflections or time reversals. Each element of the module $\mathbf{S}^\bullet$ will be disjoint union over all $P \in \mathscr{M}$, of an element from each $\mathbf{S}^\bullet(P)$. Some differentiable structure will be imposed as well. This will be defined axiomatically, shortly. An object built up of tensor products of $\mathbf{S}^\bullet$, its dual $\mathbf{S}_\bullet$ and the complex conjugates of these two spaces will be called a *spinor field* on $\mathscr{M}$. A world-vector field is thus seen to be a special case. Here, "tensor product" will be interpreted in the sense given in Section 3. The labeling system will be very useful in order to keep track of the various operations.

We shall need spinor labels of two types:

$$A, B, C, \ldots, A_0, \ldots, A_1, \ldots \tag{4.19}$$

$$A', B', C', \ldots, A_0', \ldots, A_1', \ldots \tag{4.20}$$

For each $\omega \in \mathbf{S}^\bullet$, there will correspond elements $\omega^A \in \mathbf{S}^A$, $\omega^B \in \mathbf{S}^B$, ..., $\omega^X \in \mathbf{S}^X$, ... and also elements $\bar{\omega}^{A'} \in \mathbf{S}^{A'}$, $\bar{\omega}^{B'} \in \mathbf{S}^{B'}$, ..., $\bar{\omega}^{X'} \in \mathbf{S}^{X'}$, ... As before, we can, if desired, think of ω^X as being simply the *pair* (ω, X), etc. The ring $\mathbf{S}$ will be identified as the *complex* C^∞-functions on $\mathscr{M}$:

$$\mathbf{S} = \mathbf{T} \oplus i\mathbf{T} \tag{4.21}$$

The difference between the primed and unprimed labels arises because we shall require that, whereas each of the modules $\mathbf{S}^A$, $\mathbf{S}^B$, ... over $\mathbf{S}$ is canonically isomorphic with $\mathbf{S}^\bullet$, the modules $\mathbf{S}^{A'}$, $\mathbf{S}^{B'}$, ... over $\mathbf{S}$ are each canonically *anti-isomorphic* with $\mathbf{S}^\bullet$. That is, if $\lambda, \mu \in \mathbf{S}$ and $\boldsymbol{\alpha}, \boldsymbol{\beta}, \gamma \in \mathbf{S}^\bullet$, with $\gamma = \lambda\boldsymbol{\alpha} + \mu\boldsymbol{\beta}$, then

$$\gamma^X = \lambda\alpha^X + \mu\beta^X \quad \text{and} \quad \bar{\gamma}^{X'} = \bar{\lambda}\bar{\alpha}^{X'} + \bar{\mu}\bar{\beta}^{X'} \tag{4.22}$$

We shall, in fact, regard each $\bar{\omega}^{X'}$ as being the *complex conjugate* of the corresponding ω^X and occasionally write

$$\overline{\omega^X} = \bar{\omega}^{X'}; \; \overline{\pi^{X'}} = \bar{\pi}^X \tag{4.23}$$

(We may think of the complex conjugation operation as effecting an interchange of the list (4.19) with the list (4.20), when applied to the labels themselves.) Our spinor system will then be built up just as in Section 3, but with the proviso that contractions and index substitutions can occur only between index labels of the *same* type (that is, only between primed or only between unprimed labels). Since index substitutions do not effect interchanges of primed index labels with unprimed ones, we shall regard the relative ordering of primed and unprimed indices as irrelevant. On the other hand, it will be significant to maintain an ordering (in most cases) between indices of the *same* type *even* when some are upper indices and others are lower indices. Thus, for example,

$$\rho^{AA'}{}_{B'B}{}^{Q} = \rho^{A'A}{}_{B}{}^{Q}{}_{B'} = \rho^{AA'Q}_{BB'} \neq \rho^{AA'Q}{}_{BB'} \tag{4.24}$$

Complex conjugation gives us a fifth operation on the spinor system {S}, in addition to (3.11)–(3.14):

$$\text{complex conjugation}: \mathbf{S}^{X\cdots ZW'\cdots Y'}_{P\cdots SQ'\cdots U'} \to \mathbf{S}^{W\cdots YX'\cdots Z'}_{Q\cdots UP'\cdots S'} \tag{4.25}$$

defined on each element $\rho^{X\cdots ZW'\cdots Y'}{}_{P\cdots SQ'\cdots U'}$ [expanded as in (3.9)] in the obvious way [*cf.* (4.23)].

So far, we have not incorporated into the spinor algebra, the fact that the $(t^{\mathfrak{A}}{}_{\mathfrak{B}})$ matrices have to be unimodular. This unimodular condition can be written alternatively as

$$\varepsilon^{\mathfrak{A}\mathfrak{C}} = t^{\mathfrak{A}}{}_{\mathfrak{B}}\, t^{\mathfrak{C}}{}_{\mathfrak{D}}\, \varepsilon^{\mathfrak{B}\mathfrak{D}} \qquad \text{or} \qquad \varepsilon_{\mathfrak{B}\mathfrak{D}} = t^{\mathfrak{A}}{}_{\mathfrak{B}}\, t^{\mathfrak{C}}{}_{\mathfrak{D}}\, \varepsilon_{\mathfrak{A}\mathfrak{C}} \tag{4.26}$$

Here, the Levi–Civita symbols

$$(\varepsilon_{\mathfrak{A}\mathfrak{B}}) = (\varepsilon^{\mathfrak{A}\mathfrak{B}}) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{4.27}$$

are being used. Let us denote by ε_{AB} and ε^{AB} the elements of $\mathbf{S}_{AB}$ and $\mathbf{S}^{AB}$, respectively, whose components in some basis are given by (4.27). Then (4.26) tells us that ε_{AB} and ε^{AB} are *invariant* under the active symmetry transformation of $\mathbf{S}^{\bullet}(P)$ given by (4.18). We shall regard ε_{AB}, ε^{AB} and their complex conjugates $\varepsilon_{A'B'}$, $\varepsilon^{A'B'}$ as an essential part of the internal structure of the spinor system. (It would have been more logical to write $\bar{\varepsilon}_{A'B'}$ for $\varepsilon_{A'B'}$ and $\bar{\varepsilon}^{A'B'}$ for $\varepsilon^{A'B'}$, but these bars are normally omitted.) Any basis for which the components $\varepsilon_{\mathfrak{A}\mathfrak{B}}$ have the form (4.27) will be called a normalized basis or *spin frame*.

One effect of the presence of the ε's is that they establish a canonical isomorphism between $\mathbf{S}^{\bullet}$ and its dual $\mathbf{S}_{\bullet}$. This is achieved via the lowering and raising of spinor index labels:

$$\xi^{\cdots A \cdots}_{\cdots\ \ \cdots}\,\varepsilon_{AB} = \xi^{\cdots\ \ \cdots}_{\cdots B \cdots} \qquad \varepsilon^{AB}\xi^{\cdots\ \ \cdots}_{\cdots B \cdots} = \xi^{\cdots A \cdots}_{\cdots\ \ \cdots} \tag{4.28}$$

$$\eta^{\cdots A' \cdots}_{\cdots\ \ \cdots}\,\varepsilon_{A'B'} = \eta^{\cdots\ \ \cdots}_{\cdots B' \cdots} \qquad \varepsilon^{A'B'}\eta^{\cdots\ \ \cdots}_{\cdots B' \cdots} = \xi^{\cdots A' \cdots}_{\cdots\ \ \cdots} \tag{4.29}$$

(In view of this, it may be felt that it is redundant to use *both* lower and upper labels. The reason for maintaining the distinction is essentially a matter of "bookkeeping." In fact, a notation *is* possible which does *not* maintain this distinction, but it looks a little strange since spin vectors must then anticommute! I have chosen to be conventional in this respect.) Note that

$$\xi^{\cdots}_{\cdots X}{}^{\cdots X \cdots}_{\cdots} = -\xi^{\cdots X \cdots}_{\cdots}{}^{\cdots}_{\cdots X \cdots} \tag{4.30}$$

(and similarly for primed indices) because of the antisymmetry:

$$\varepsilon_{AB} = -\varepsilon_{BA} \qquad \varepsilon^{AB} = -\varepsilon^{BA} \tag{4.31}$$

so we must be careful about orderings when we raise and lower indices. For this reason, and so as to avoid notational confusion at a later stage, the symbols δ^A_B and $\delta^{A'}_{B'}$, corresponding to (3.18) will *not* be used here. Instead, the equivalent symbols $\varepsilon_B{}^A$, $\varepsilon_{B'}{}^{A'}$ will be employed. We have

$$\varepsilon_{AB}\varepsilon^{CB} = \varepsilon_A{}^C = -\varepsilon^C{}_A \qquad \varepsilon_A{}^A = 2 \tag{4.32}$$

$$\xi^{\cdots}_{\cdots B}{}^{\cdots}_{\cdots} = \xi^{\cdots}_{\cdots A}{}^{\cdots}_{\cdots}\varepsilon_B{}^A \qquad \xi^{\cdots B \cdots}_{\cdots}{}^{\cdots}_{\cdots} = \xi^{\cdots A \cdots}_{\cdots}{}^{\cdots}_{\cdots}\varepsilon_A{}^B \tag{4.33}$$

so that $\varepsilon_B{}^A$ has the properties (3.19) required of a "δ_B." (Note that the raising and lowering operations are consistent.) With any choice of spin frame $\varepsilon_{\mathfrak{A}}{}^A$, $\varepsilon_A{}^{\mathfrak{A}}$ we have

$$\left(\varepsilon_{\mathfrak{A}}{}^{\mathfrak{B}}\right) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4.34}$$

If a spinor is skew-symmetrical in three or more indices then it vanishes (from two-dimensionality). Hence,

$$\varepsilon_{AB}\varepsilon_{CD} + \varepsilon_{AC}\varepsilon_{DB} + \varepsilon_{AD}\varepsilon_{BC} = 0 \tag{4.35}$$

One consequence of (4.35) is that if $\zeta_{\dots A \dots B \dots}$ is skew-symmetrical in A and B:

$$\zeta_{\dots A \dots B \dots} = -\zeta_{\dots B \dots A \dots} \tag{4.36}$$

then

$$\zeta_{\dots A \dots B \dots} = \tfrac{1}{2}\varepsilon_{AB}\zeta_{\dots X \dots}{}^{X}{}_{\dots} \tag{4.37}$$

This property is closely related to the representation theory of $SL(2, \mathbf{C})$, since it indicates that it is only spinors totally symmetrical in all unprimed indices and in all primed indices which cannot be reduced into simpler parts. The components of a spinor with r symmetric lower unprimed indices and with s symmetric lower primed indices are said to transform according to the $D(r/2,s/2)$ irreducible representation of the Lorentz group.

Let us now incorporate the world-tensors into our spinor scheme, introducing world-tensor labels (3.1) by identifying them as *pairs* of spinor labels as follows:

$$a = AA', b = BB', c = CC', \dots, a_0 = A_0 A_0', \dots \tag{4.38}$$

This gives us the subsystem ($\mathbf{S}$, $\mathbf{S}^a$, $\mathbf{S}^b$, $\dots$, $\mathbf{S}_a$, $\dots$, $\mathbf{S}^{x \cdots z}_{u \cdots w}$, $\dots$) of *complex world tensors*, closed under the operations (3.11)–(3.14) and also under the complex conjugation operation (4.25), which here acts: $\mathbf{S}^{x \cdots z}_{u \cdots w} \rightarrow \mathbf{S}^{x \cdots z}_{u \cdots w}$.

The elements which are invariant under complex conjugation are the *real* world-tensors: $\bar{\alpha}_{u\cdots w}^{x\cdots z} = \alpha_{u\cdots w}^{x\cdots z}$. The real world-tensor fields from a subsystem $\{\mathrm{T}\}$ of $\{\mathrm{S}\}$:

$$\{\mathrm{T}\} = (\mathrm{T}, \mathrm{T}^{a}, \mathrm{T}^{b}, \ldots, \mathrm{T}_{a}, \ldots, \mathrm{T}_{u\cdots w}^{x\cdots z}, \ldots) \tag{4.39}$$

To see that this process is really what is involved in (4.13), let us consider a spin frame $\varepsilon_{\mathfrak{A}}{}^{A}$, $\varepsilon_{A}{}^{\mathfrak{A}}$ for $\mathrm{S}^{\bullet}$ and define (with $\varepsilon_{A'}{}^{\mathfrak{A}'} = \overline{\varepsilon_{A}{}^{\mathfrak{A}}}$)

$$\begin{aligned}
\delta_{a}^{0} &= \frac{1}{\sqrt{2}}(\varepsilon_{A}{}^{0}\varepsilon_{A'}{}^{0'} + \varepsilon_{A}{}^{1}\varepsilon_{A'}{}^{1'}) \\
\delta_{a}^{1} &= \frac{1}{\sqrt{2}}(\varepsilon_{A}{}^{0}\varepsilon_{A'}{}^{0'} - \varepsilon_{A}{}^{1}\varepsilon_{A'}{}^{1'}) \\
\delta_{a}^{2} &= \frac{1}{\sqrt{2}}(\varepsilon_{A}{}^{1}\varepsilon_{A'}{}^{0'} + \varepsilon_{A}{}^{0}\varepsilon_{A'}{}^{1'}) \\
\delta_{a}^{3} &= \frac{i}{\sqrt{2}}(\varepsilon_{A}{}^{1}\varepsilon_{A'}{}^{0'} - \varepsilon_{A}{}^{0}\varepsilon_{A'}{}^{1'})
\end{aligned} \tag{4.40}$$

This is readily seen to give (4.13), where $u^{a} = u^{AA'}$ and

$$u^{\mathfrak{a}} = u^{a}\delta_{a}^{\mathfrak{a}} \qquad u^{\mathfrak{A}\mathfrak{A}'} = u^{AA'}\varepsilon_{A}{}^{\mathfrak{A}}\varepsilon_{A'}{}^{\mathfrak{A}'} \tag{4.41}$$

as is required by the convention (3.20). The definitions (4.40) give us a standard way of obtaining a world-vector basis from any spin frame. However, it is important that we not be tied down to this particular correspondence in all cases. Sometimes we shall prefer to obtain our world-vector basis as the one naturally obtained from a *coordinate system*. In other circumstances the *null-tetrad*

$$\delta_{0}^{a} = \varepsilon_{0}{}^{A}\varepsilon_{0'}{}^{A'} \qquad \delta_{1}^{a} = \varepsilon_{1}{}^{A}\varepsilon_{1'}{}^{A'} \qquad \delta_{2}^{a} = \varepsilon_{0}{}^{A}\varepsilon_{1'}{}^{A'} \qquad \delta_{3}^{a} = \varepsilon_{1}{}^{A}\varepsilon_{0'}{}^{A'} \tag{4.42}$$

is a convenient basis to use. In each case, we can define the Infeld–van der Waerden translation symbols [46, 110]:

$$\sigma_{\mathfrak{a}}^{\mathfrak{A}\mathfrak{A}'} = \delta_{\mathfrak{a}}^{a}\varepsilon_{A}{}^{\mathfrak{A}}\varepsilon_{A'}{}^{\mathfrak{A}'} \qquad \sigma_{\mathfrak{A}\mathfrak{A}'}^{\mathfrak{a}} = \delta_{a}^{\mathfrak{a}}\varepsilon_{\mathfrak{A}}{}^{A}\varepsilon_{\mathfrak{A}'}{}^{A'} \tag{4.43}$$

(Recall that $a = AA'$, so a contraction is taking place between the *abstract* labels. On the other hand, no connection is to be implied between $\mathfrak{a}$, and $\mathfrak{A}$, $\mathfrak{A}'$, so no summation is involved. Each of (4.43) stands for 16 equations.) In fact, the Infeld–van der Waerden symbols are really just δ_{a}^{b}, but with the upper and lower indices expressed according to different kinds of bases. The relation between spinor components and world-tensor components is then exemplified by

$$\psi_{\mathfrak{a}}{}^{\mathfrak{b}} = \psi_{\mathfrak{A}\mathfrak{A}'}{}^{\mathfrak{B}\mathfrak{B}'}\sigma_{\mathfrak{a}}^{\mathfrak{A}\mathfrak{A}'}\sigma_{\mathfrak{B}\mathfrak{B}'}^{\mathfrak{b}} \qquad \psi_{\mathfrak{A}\mathfrak{A}'}{}^{\mathfrak{B}\mathfrak{B}'} = \psi_{\mathfrak{a}}{}^{\mathfrak{b}}\sigma_{\mathfrak{A}\mathfrak{A}'}^{\mathfrak{a}}\sigma_{\mathfrak{b}}^{\mathfrak{B}\mathfrak{B}'} \tag{4.44}$$

Now, consider twice the determinant of (4.13):

$$u^{\mathfrak{A}\mathfrak{A}'}u^{\mathfrak{B}\mathfrak{B}'}\varepsilon_{\mathfrak{A}\mathfrak{B}}\varepsilon_{\mathfrak{A}'\mathfrak{B}'} = u^{\mathfrak{a}}u^{\mathfrak{b}}g_{\mathfrak{a}\mathfrak{b}} \tag{4.45}$$

CHAPTER 36

This becomes

$$u^a u^b(\varepsilon_{AB}\varepsilon_{A'B'} - g_{ab}) = 0 \tag{4.46}$$

on translation back to the abstract labels [*cf.* (3.21)], so [because of the symmetry of g_{ab} and of $\varepsilon_{AB}\varepsilon_{A'B'}$ under interchange of a, b and the fact that (4.46) holds for all u^a] we get

$$g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'} \tag{4.47}$$

Equation (4.47) expresses the fundamental way in which the metric of space-time is defined by its spinor structure. Note, in particular, that the $(+, -, -, -)$ *signature is built into the formalism* by the identification (4.47), because with the choice of world-vector basis (4.40), we get (4.15). As a consequence of (4.47), (4.32), (4.31) we have

$$g^{ab} = \varepsilon^{AB}\varepsilon^{A'B'} \qquad \delta_a^b = \varepsilon_A{}^B\varepsilon_{A'}{}^{B'} \qquad g_{ab} = g_{ba} \qquad g_{ab}g^{bc} = \delta_a^c \tag{4.48}$$

so the spinor raising and lowering conventions yield

$$\zeta^{\cdots}_{\cdots a \cdots}{}^{\cdots} = \zeta^{\cdots b \cdots}_{\cdots \quad \cdots} g_{ab} \qquad \zeta^{\cdots b \cdots}_{\cdots \quad \cdots} = \zeta^{\cdots}_{\cdots a \cdots}{}^{\cdots} g^{ab} \tag{4.49}$$

Thus, the notation is consistent with the usual conventions.

So far, nothing has been said to relate the spinor structure to the manifold $\mathscr{M}$ as a whole. In order to connect the various $\mathbf{S}^{\bullet}(P)$'s to each other, it will be necessary to consider differentiation. Let us define axiomatically the various rules we shall need. We require an operation of *covariant differentiation*:

$$\nabla_{XY'}: \mathbf{S}_{F \cdots MH' \cdots K}^{P \cdots SR' \cdots T'} \to \mathbf{S}_{F \cdots MXH' \cdots K'Y'}^{P \cdots SR' \cdots T'} \tag{4.50}$$

satisfying

$$\nabla_{XY'}(\psi^{\cdots}_{\cdots} + \chi^{\cdots}_{\cdots}) = \nabla_{XY'}\psi^{\cdots}_{\cdots} + \nabla_{XY'}\chi^{\cdots}_{\cdots} \tag{4.51}$$

$$\nabla_{XY'}(\psi^{\cdots}_{\cdots}\chi^{\cdots}_{\cdots}) = \chi^{\cdots}_{\cdots}\nabla_{XY'}\psi^{\cdots}_{\cdots} + \psi^{\cdots}_{\cdots}\nabla_{XY'}\chi^{\cdots}_{\cdots} \tag{4.52}$$

$$\nabla_{XY'}\varepsilon_{AB} = 0$$

$$\nabla_{XY'}\varepsilon^{AB} = 0 \tag{4.53}$$

$$\psi^{\cdots}_{\cdots} = \nabla_{XY'}\chi^{\cdots}_{\cdots} \qquad \text{IMPLIES} \qquad \overline{\psi^{\cdots}_{\cdots}} = \nabla_{YX'}\overline{\chi^{\cdots}_{\cdots}} \tag{4.54}$$

$$\psi^{\cdots A \cdots\cdots}_{\cdots\cdots B \cdots} = \nabla_{XY'}\chi^{\cdots A \cdots\cdots}_{\cdots\cdots B \cdots} \qquad \text{IMPLIES} \qquad \psi^{\cdots A \cdots\cdots}_{\cdots\cdots A \cdots} = \nabla_{XY'}\chi^{\cdots A \cdots\cdots}_{\cdots\cdots A \cdots} \tag{4.55}$$

$\nabla_{XY'}$ COMMUTES WITH ANY INDEX SUBSTITUTION (NOT INVOLVING X OR Y') (4.56)

$$\nabla_x\nabla_z\phi = \nabla_z\nabla_x\phi \tag{4.57}$$

FOR ANY DERIVATION ξ ON $\mathbf{T}$, THERE EXISTS A UNIQUE ELEMENT $\xi^a \in \mathbf{T}^a$ SUCH THAT $\xi(\phi) = \xi^a\nabla_a\phi$, FOR EACH $\phi \in \mathbf{T}$ (4.58)

Actually, the axioms (4.51)–(4.58) are not quite independent of one another, but they express all the properties that we require. Note, that it is (4.58) which connects {S} with the tangent bundle of $\mathcal{M}$. The world-vectors become tangent vectors to $\mathcal{M}$. The operator $\nabla_{XY'}$ is uniquely defined by all these properties. (Clearly, for any other pair of labels M, N', say, we can also define $\nabla_{MN'}$ by index substitution.) Conversely, we may ask whether the existence of a system {S} satisfying all the requirements we have imposed implies any restriction on M as a differentiable manifold, or as a pseudo-Riemannian manifold with given $(+, -, -, -)$ metric (with which the g_{ab} of (4.47) is supposed to agree). The answer is that *locally* no such restriction is implied (that is an {S} exists for some open set containing any given $P \in \mathcal{M}$) whereas *globally* there *are* such restrictions. A similar remark applies to the uniqueness of {S}. These global restrictions on $\mathcal{M}$ will enter into the discussion in the next section.

5 THE INTERPRETATION OF A SPIN-VECTOR

In the previous section, spinors were introduced completely formally. Only the elements of {T} were given a direct interpretation in space-time terms. One might imagine that any geometrical interpretation of the remaining elements of {S} would necessarily be very indirect. However, this turns out not to be the case. Any spin-vector has, in fact, a very graphic space-time interpretation up to sign. Even the sign can be given a meaning in terms of physical constructions.

Let us try to interpret the spin-vector

$$\omega^A \in S^A(P)$$

The most obvious world-tensor we can build from ω^A is the world-vector

$$u^a = \omega^A \bar{\omega}^{A'} \tag{5.1}$$

Now $u^a u_a = |\omega^A \omega_A|^2 = 0$ (since $\omega^A \omega_A = -\omega_A \omega^A = 0$), so u^a points along the null cone at P. In fact, *every* real null world-vector has either the form (5.1) or the form

$$v^a = -\omega^A \bar{\omega}^{A'} \tag{5.2}$$

and every complex world-vector, null in the sense $w^a w_a = 0$, has the form

$$w^a = \omega^A \pi^{A'} \tag{5.3}$$

These facts follow at once from the representation (4.13), since a (2×2) matrix of vanishing determinant has rank at most unity and is therefore the outer product of two vectors. Note that the existence of the spinor system {S} gives us an absolute distinction between the two null half-cones at P. We can *define* the *future-pointing* null vectors to be those for which (5.1) holds (with $\omega^A \neq 0$) and the *past-pointing* null vectors to be those for which (5.2) holds ($\omega^A \neq 0$). (This is consistent with (4.13) if positive u^0 corresponds to time increasing.) This gives us the first global restriction on $\mathcal{M}$, as a pseudo-Riemannian manifold, implied by the existence of {S}: $\mathcal{M}$ must be *time-orientable*. That is to say, the division of the null half-cones of $\mathcal{M}$ into two classes, "future" and "past," can be made continuously over the whole manifold.

Equation (5.1) tells us, then, that a future-pointing real null world-vector is uniquely defined by any nonzero spin-vector. However, many spin-vectors will correspond to the same null world-vector since (5.1) is invariant under

$$\omega^A \to e^{i\theta}\omega^A \tag{5.4}$$

(θ real). To interpret the "phase" of ω^A we have to try something different. Let us "square" ω^A instead: $\omega^A\omega^B$. Then, to introduce as many primed indices as unprimed, we multiply by $\varepsilon^{A'B'}$. Finally, to get a real world-tensor we add the complex conjugate:

$$p^{ab} = \omega^A\omega^B\varepsilon^{A'B'} + \varepsilon^{AB}\bar{\omega}^{A'}\bar{\omega}^{B'} \tag{5.5}$$

Now p^{ab} is readily seen to be skew and simple:

$$p^{ab} = u^a k^b - k^a u^b \tag{5.6}$$

where

$$k^a = \omega^A\bar{\kappa}^{A'} + \kappa^A\bar{\omega}^{A'} \tag{5.7}$$

κ^A being any element of $S^A(p)$ subject to $\omega_A\kappa^A = 1$, whence (by (4.37)) $\omega^A\kappa^B - \kappa^A\omega^B = \varepsilon^{AB}$. The vector k^a is a real, spacelike of length $\sqrt{2}$ and orthogonal to u^a:

$$\bar{k}^a = k^a \qquad k^a k_a = -2 \qquad k^a u_a = 0 \tag{5.8}$$

It is defined by p^{ab} up to the addition of a (real) multiple of u^a. The directions k^a therefore define a half-plane element at P (two-dimensional) which is tangent to the null cone along the null direction of u^a. Thus ω^A defines a kind of *null flag* [70, 73, 114] in the tangent space at P, whose flagpole is the future-pointing null vector u^a and whose flag plane is this null 2-plane through u^a (Fig. 7).

To get a better picture of the nature of this flag plane, let us take a section in the tangent space at P by a spacelike hyperplane not through P. This cuts the cone in a sphere S^2. The flagpole intersects S^2 in a *point* U and the flag plane gives us a *unit tangent vector* to S^2 at U (which we may identify with that choice of $2^{-1/2}k^a$ which lies parallel to our hyperplane). We can get a direct physical interpretation of this S^2 as the "celestial sphere"

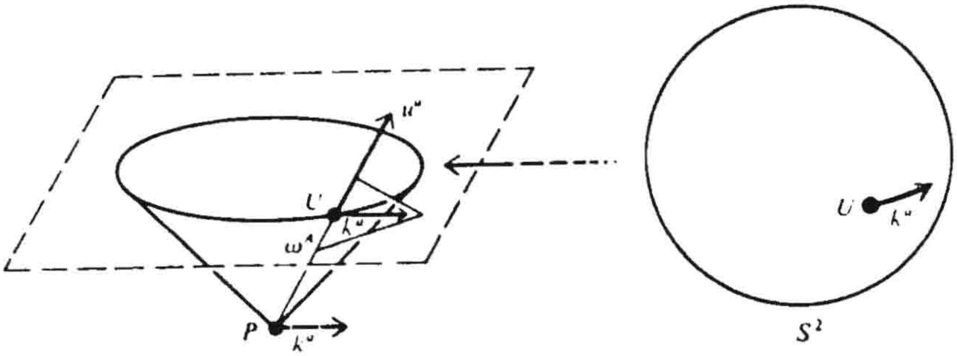


FIGURE 7. The spin vector ω^A at P defines a null flag. This may be pictured as a tangent vector at a point U on the celestial sphere.

(or field of vision) of an observer at P. (We can choose our spacelike hyperplane orthogonal to the observer's world line at P.) A photon arriving at P has a world line (a null geodesic) which has a null tangent at P and is therefore uniquely associated with a point U of S^2. We may also describe the *polarization* of the photon in terms of a flag plane, that is, as a tangent vector at U to the celestial sphere. Thus, a spin vector emerges as a natural object for the description of a polarized photon. This relation between spinors and zero rest-mass particle fields (of various spins) will be discussed in more detail in Section 8.

One final remark concerning the structure of the celestial sphere is appropriate here. The Riemannian metric induced on S^2 by the above construction will depend on our choice of spacelike hyperplane, but the *conformal structure* of S^2 is, on the other hand, well-defined. One way of seeing this is to refer to (4.18), which when applied to $\zeta = \omega^0/\omega^1$ yields the fractional linear transformation $\zeta \rightarrow (t^0{}_0\zeta + t^0{}_1)/(t^1{}_0\zeta + t^1{}_1)$. The complex number ζ defines the point U, the complex ζ-plane being actually related to S^2 by a stereographic projection [71]. Both these transformations are conformal (and, in fact, send circles into circles). Since (4.18) corresponds to a Lorentz transformation, it follows that the celestial spheres of two observers at P are related to each other by a conformal transformation [71, 106] (and a circular pattern seen by one observer will also appear as circular to the other). Another way of invariantly obtaining the conformal structure of S^2 will emerge in a moment.

If we multiply ω^A by a nonzero complex number $\lambda = re^{i\theta}$ (r, θ real), then u^a gets multiplied by r^2 and (taking $\kappa^A \rightarrow \lambda^{-1}\kappa^A$) k^a gets rotated through an angle 2θ (in the plane spanned by k^a and $h^a = i\omega^A\bar{\kappa}^{A'} - i\kappa^A\bar{\omega}^{A'}$). Thus, we see that *under* $\omega^A \rightarrow \lambda\omega^A$, *the extent of the flagpole is multiplied by a factor* $\lambda\bar{\lambda}$ *and the flag plane rotates through an angle* $2 \arg \lambda$. Note that this gives a direct way of defining the conformal structure of our S^2 in terms of spinors. If U is a point of S^2, then the angle between the two tangent vectors to S^2

at U given by ω^A and $e^{i\theta}\omega^A$, respectively, is simply 2θ. Having an invariant definition of angles on S^2 we therefore have an invariantly defined conformal structure for S^2. Note also that we have an invariantly defined *orientation* for S^2 and therefore also for $\mathcal{M}$. We can *define* the notion of *right-handedness* by simply specifying that "right-handed" is the sense in which the flag plane of $e^{i\theta}\omega^A$ rotates as θ increases. (This is consistent with the reference frame referred to in (4.13) being a right-hand one.) This gives us our second global restriction on $\mathcal{M}$ as a pseudo-Riemannian manifold: $\mathcal{M}$ must, in addition to being time-orientable, also be *space-orientable*. As a topological manifold, therefore, $\mathcal{M}$ is orientable.

The third and final global restriction on $\mathcal{M}$ as a pseudo-Riemannian (or topological) manifold, which is implied by the presence of $\{S\}$, arises in relation to another feature of the above representation of a spin-vector as a null flag. We note that both the u^a of (5.1) and the p^{ab} of (5.6) are invariant under

$$\omega^A \to -\omega^A \tag{5.9}$$

Explicitly, if we consider the flag plane given by $e^{i\theta}\omega^A$ as θ varies from 0 to π, we see that it executes one complete rotation through 2π and returns to its starting position, whereas ω^A becomes replaced by its negative. Only after a further complete rotation of the flag plane through 2π will the spin vector return to its original value. It is clear, moreover, that no additional local geometric structure added to the null flag, can distinguish ω^A from $-\omega^A$ (interpreting "local geometric structure" to refer to structure in the tangent space to $\mathcal{M}$ at P). For, a rotation of any such structure through 2π would return to its original configuration, whereas ω^A would undergo (5.9).

On the other hand, a *nonlocal* geometrical distinction between ω^A and $-\omega^A$ is possible. To obtain this, we consider a particular null flag at a particular point 0 of $\mathcal{M}$. Call this the *standard flag*. For simplicity, let us restrict our attention to a simply-connected region of $\mathcal{M}$ (or else work with the universal covering manifold of $\mathcal{M}$). The idea, then, is to identify a spin vector, not just with a null flag, but as an *equivalence class of flag paths*.[18] A flag path is a continuous sequence of null flags on $\mathcal{M}$ which start with the standard flag and end with some null flag at a point P of $\mathcal{M}$ (Fig. 8). Two flag paths are equivalent if they can be continuously distorted one into the other, keeping the end-points fixed, and the null flags at the end-points fixed. The existence of spin vectors now depends on the existence of *two* equivalence classes corresponding to each null flag at P. If we consider flag paths based on a *fixed* curve on $\mathcal{M}$ and ending at a fixed null flag, then it follows from the

---

[18] I have been deliberately somewhat vague as to whether the geometrical structure corresponding to the *zero* spin vector at P should be called a "null flag." However, in the case of a *flag path*, the "zero null flags" are to be definitely excluded.

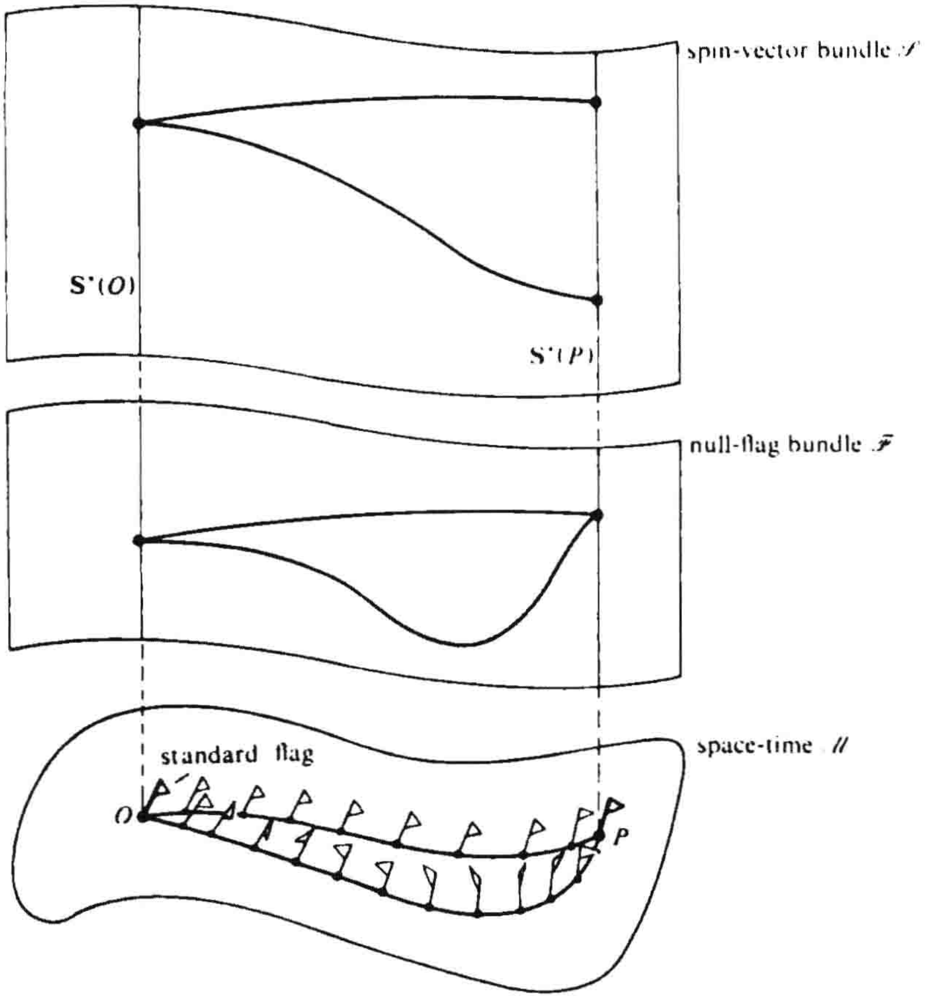

spin-vector bundle $\mathscr{S}$
S'(O)
S'(P)
null-flag bundle $\mathscr{F}$
space-time $\mathscr{M}$
standard flag
O
P

FIGURE 8. Flag paths in $\mathscr{M}$ and their images in $\mathscr{F}$ and $\mathscr{S}$.

topology of the Lorentz group that there are precisely two equivalence classes of flag paths associated with the curve, which are related to each other by a net relative rotation through 2π between 0 and P. But if we allow the curve, on which the flag path is based, to vary over the manifold and then return to its original position, it may be that these equivalence classes become united inside a single equivalence class. Then spin-vectors could not be defined on the manifold. Another way of stating this property arises if we allow P to coincide with 0. If the (closed) flag path consisting simply of a rotation through 2π of the standard flag about its flagpole can be continuously distorted, by moving the curve about the manifold, into the constant flag path (that is, no motion of the null flag at all), then a spinor structure does not exist.

The above construction clearly actually refers to the null-flag bundle $\mathscr{F}$ of $\mathscr{M}$ (an 8-dimensional space of all null flags at all points of $\mathscr{M}$). A flag path corresponds to a curve in $\mathscr{F}$. If $\mathscr{M}$ is simply-connected, the above construction gives the spin-vector bundle $\mathscr{S}$ (space of all spin vectors at all points of $\mathscr{M}$) as the universal (twofold) covering space of $\mathscr{F}$—*provided* that $\mathscr{F}$ possesses a twofold covering. If $\mathscr{F}$ does not possess a twofold covering (that is, is simply-connected) then $\{S\}$ does not exist for $\mathscr{M}$. When $\{S\}$ does exist for $\mathscr{M}$, then we can interpret the elements of $S^{\bullet}$ as the C^{∞} cross sections of $\mathscr{S}$. When $\mathscr{M}$ itself is *not* simply-connected the situation becomes more complicated because flag paths belonging to nonhomotopic curves on $\mathscr{M}$ cannot be compared. The spinor structure $\{S\}$ for $\mathscr{M}$, if it exists, becomes *nonunique*. But in any case, for the *existence* of a spinor structure we can refer to the universal covering manifold of $\mathscr{M}$ and the situation is the same as before.

The foregoing discussion has been given in terms of the bundle $\mathscr{F}$ of null-flags since this seems to be the most direct geometrical route to the space-time description of a spin-vector. It is more usual to use the bundle of all orthonormal frames on $\mathscr{M}$, however, and this is more appropriate for the general n-dimensional discussion. The two procedures are readily seen to be equivalent for a space-time manifold. Also, since $\mathscr{M}$ is orientable (that is vanishing of Steifel–Whitney class w_1), it is not difficult to adapt the above arguments to show that the remaining topological condition on $\mathscr{M}$ implied by the existence of $\{S\}$ is that its *Steifel–Whitney class* w_2 should vanish [38, 57]. This means that for every 2-surface in M, we can put a set of three linearly independent vectors at each point of the surface in a continuous way. (See [113a] and Chs. V and XVII by Lichnerowicz and Bott for discussions of this.) It is of interest that "plausible" space-time manifolds $\mathscr{M}$ do exist[19] which are space- and time-orientable and which contain no closed (or almost closed) timelike curves, but for which $w_2 \neq 0$. For such $\mathscr{M}$, the system $\{S\}$ could not exist.

We may ask what are the physical grounds for believing in the existence of a spinor structure for space-time. How much physical reality can one indeed assign to an object which is not returned to its original state when a rotation through 360° is applied to it? The idea seems totally foreign to our experiences. Nevertheless, the wave functions of electrons, protons, neutrons, neutrinos, and many other particles do in fact behave in this way. But we might well argue that this is not conclusive, because a wave function is a somewhat nebulous concept. In any case, surely it is only the overall sign of the wave function which is altered upon rotation through 2π and is not the sign of a wave function supposed to be unobservable? In an interesting recent article, Aharonov and Susskind [1] argue that this view cannot be maintained. They show how it is possible, in principle, to construct an apparatus which, depending precisely on this spinor nature of the electron wave equation, exhibits such behavior under rotation on a *macroscopic* scale. The apparatus consists of two parts which, when fitted together, result in a detectable electric current flowing from one part to the other. The parts are then separated and one is rotated through 2π relative to the other. They are again fitted together in precisely the same relative orientation as before. A detectable current again flows from one part to the other, but now the direction of flow is reversed! Only by separating the parts and applying a further rotation, through 2π, of one part relative to the other, is the original direction of flow restored. Thus we have a situation in which the geometry of the apparatus, in the usual sense, does not define its behavior. In this case, by widening the concept of geometry to include spinorial quantities, we can again think of the geometrical configuration as determining the behavior.

---

[19] Take a complex projective plane $\mathscr{P}$ (a real 4-manifold) and choose a (positive definite) C^{∞} Riemannian metric on $\mathscr{P}$. Let t be a real C^{∞} function on $\mathscr{P}$ with only isolated critical points. Remove these points from the manifold and use the unit vector corresponding to the gradient of t to construct a C^{∞} Lorentzian $(+, -, -, -)$ pseudo-Riemannian metric. The resulting "space-time" is time-orientable, space-orientable without closed (that is, $\cong S^1$) timelike curves [in fact, satisfying strong causality (11.1)], but possesses no spin structure ($w_2 \neq 0$). I am grateful to R. Bott and S. Smale for this example.

Added in proof: In a recent article (submitted to *J. Math. Phys.*), Geroch has shown that for a non-compact space-time, existence of a spinor structure is equivalent to global existence of a continuous field of orthonormal tetrads. He also shows that our assumption (4.5) is, in fact, redundant.

The Aharonov-Susskind apparatus gives us a "physical" way of picturing a spin-vector. We may think of one part of the apparatus as being attached to our standard flag and allow the other part to be attached to another null flag which moves about the space-time. Then the apparatus keeps track of the parity of the number of complete relative rotations which take place, so in effect we have a spin-vector. If it could be convincingly argued that such pieces of apparatus could (in principle) retain their "memory" over large (say cosmological) distances and times, then we should have a fairly conclusive argument that w_2 for the universe really does vanish!

6 EXPLICIT CURVATURE FORMULAS

The notation of Section 3 yields, in an automatic way, all the familiar basic formulas involving the curvature, Christoffel symbols, and Ricci rotation coefficients, once certain basic definitions are given. This will be indicated here, in conjunction with a parallel description of the various spinor formulas, based on the notation of Section 4.

Consider a basis $\delta^{a}_{\mathbf{a}}$ for $\mathbf{T}^{a}$ and a (possibly unrelated) basis $\varepsilon_{\mathfrak{A}}{}^{A}$ for $\mathbf{S}^{A}$. Let $\delta^{\mathbf{a}}_{a}$, $\varepsilon_{A}{}^{\mathfrak{A}}$ be the corresponding dual bases. Define

$$\Gamma^{\mathbf{a}}{}_{\mathbf{bc}} = \delta^{\mathbf{a}}_{a} \nabla_{\mathbf{c}} \delta^{a}_{\mathbf{b}} = -\delta^{a}_{\mathbf{b}} \nabla_{\mathbf{c}} \delta^{\mathbf{a}}_{a} \tag{6.1}$$

and

$$\gamma_{\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}'} = \varepsilon_{A\mathfrak{B}} \nabla_{\mathfrak{C}\mathfrak{D}'} \varepsilon_{\mathfrak{A}}{}^{A} = \varepsilon_{\mathfrak{B}}{}^{A} \nabla_{\mathfrak{C}\mathfrak{D}'} \varepsilon_{\mathfrak{A}A} \tag{6.2}$$

[Here $\nabla_{\mathfrak{c}} \equiv \delta^{c}_{\mathfrak{c}} \nabla_c$ and $\nabla_{\mathfrak{C}\mathfrak{D}'} \equiv \varepsilon_{\mathfrak{C}}{}^{C} \varepsilon_{\mathfrak{D}'}{}^{D'} \nabla_{CD'}$, in accordance with (3.20). The equivalence of the two formulas in (6.1) and in (6.2) is an immediate consequence of the rules (4.51)–(4.56), (3.17), (4.30), (4.31).] The quantities $\Gamma^{\mathfrak{a}}{}_{\mathfrak{b}\mathfrak{c}}$ specialize either to *Ricci rotation coefficients* [in the case when the basis frame is chosen so that $g_{\mathfrak{a}\mathfrak{b}}$ has a specific form, for example (4.15)] or to *Christoffel symbols* (when the basis is that naturally derived from a coordinate system). The quantities $\gamma_{\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}'}$, are called *spin coefficients* [64]. As an example, let us derive the usual Christoffel symbol symmetry. Let $x^{0}, \ldots, x^{3}$ be a local coordinate system on $\mathcal{M}$. Then the associated coordinate basis is given by

$$\delta^{\mathfrak{a}}_{a} = \nabla_{a} x^{\mathfrak{a}} \tag{6.3}$$

with $\delta^{a}_{\mathfrak{a}}$ dual to this. Thus we have, in this case,

$$\Gamma^{\mathfrak{a}}{}_{\mathfrak{b}\mathfrak{c}} = -\delta^{a}_{\mathfrak{b}} \nabla_{\mathfrak{c}} \nabla_{a} x^{\mathfrak{a}} = -\delta^{d}_{\mathfrak{c}} \delta^{a}_{\mathfrak{b}} \nabla_{d} \nabla_{a} x^{\mathfrak{a}} \tag{6.4}$$

which is manifestly symmetric in $\mathfrak{c}\mathfrak{b}$ because of the vanishing torsion condition (4.57). Also the spin coefficient symmetry

$$\gamma_{\mathfrak{A}\mathfrak{B}\mathfrak{C}\mathfrak{D}'} = \gamma_{\mathfrak{B}\mathfrak{A}\mathfrak{C}\mathfrak{D}'} \tag{6.5}$$

is immediate from (6.2).

The formulas (6.1) and (6.2), together with (3.20), (3.21) yield the usual formulas for covariant differentiation in terms of components. This is exemplified by

$$\begin{aligned} \delta^{c}_{\mathfrak{c}} \delta^{b}_{\mathfrak{b}} \delta^{\mathfrak{a}}_{a} (\nabla_{c} \xi^{a}{}_{b}) &= \delta^{b}_{\mathfrak{b}} \delta^{\mathfrak{a}}_{a} \nabla_{\mathfrak{c}} (\xi^{\mathfrak{d}}{}_{\mathfrak{e}} \delta^{a}_{\mathfrak{d}} \delta^{\mathfrak{e}}_{b}) \\ &= \nabla_{\mathfrak{c}} \xi^{\mathfrak{a}}{}_{\mathfrak{b}} + \xi^{\mathfrak{d}}{}_{\mathfrak{b}} \Gamma^{\mathfrak{a}}{}_{\mathfrak{d}\mathfrak{c}} - \xi^{\mathfrak{a}}{}_{\mathfrak{e}} \Gamma^{\mathfrak{e}}{}_{\mathfrak{b}\mathfrak{c}} \end{aligned} \tag{6.6}$$

$$\begin{aligned} \varepsilon_{\mathfrak{C}}{}^{C} \varepsilon_{\mathfrak{D}'}{}^{D'} \varepsilon_{\mathfrak{A}}{}^{A} \varepsilon_{B'}{}^{\mathfrak{B}'} (\nabla_{CD'} \theta_{A}{}^{B'}) &= \varepsilon_{\mathfrak{A}}{}^{A} \varepsilon_{B'}{}^{\mathfrak{B}'} \nabla_{\mathfrak{C}\mathfrak{D}'} (\theta_{\mathfrak{E}}{}^{\mathfrak{F}'} \varepsilon_{A}{}^{\mathfrak{E}} \varepsilon_{\mathfrak{F}'}{}^{B'}) \\ &= \nabla_{\mathfrak{C}\mathfrak{D}'} \theta_{\mathfrak{A}}{}^{\mathfrak{B}'} - \theta_{\mathfrak{E}}{}^{\mathfrak{B}'} \Gamma^{\mathfrak{E}}{}_{\mathfrak{A}\mathfrak{C}\mathfrak{D}'} + \theta_{\mathfrak{A}}{}^{\mathfrak{F}'} \Gamma^{\mathfrak{B}'}{}_{\mathfrak{F}'\mathfrak{D}'\mathfrak{C}} \;. \end{aligned} \tag{6.7}$$

(Recall that when acting on quantities without abstract labels, ∇.. is just the ordinary gradient operator.)

Let us now define the *curvature tensor* and *curvature spinors*. Put

$$R^{a}{}_{bcd} = 2\delta^{\mathfrak{a}}_{a} \nabla_{[c} \nabla_{d]} \delta^{a}_{\mathfrak{b}} \tag{6.8}$$

and, defining

$$\Box_{AB} \equiv \nabla_{Y'(A} \nabla^{Y'}_{B)} \qquad \Box_{A'B'} \equiv \nabla_{Y(A'} \nabla^{Y}_{B')} \tag{6.9}$$

set

$$\Psi_{ABCD} = \varepsilon_{D\mathfrak{D}} \Box_{(AB} \varepsilon_{C)}{}^{\mathfrak{D}} \tag{6.10}$$

$$\Phi_{CDA'B'} = \varepsilon_{D\mathfrak{D}} \Box_{A'B'} \varepsilon_{C}{}^{\mathfrak{D}} \tag{6.11}$$

$$\Lambda = \tfrac{1}{6} \varepsilon_{A\mathfrak{D}} \Box^{AB} \varepsilon_{B}{}^{\mathfrak{D}} \tag{6.12}$$

(Here the conventions are being employed that square brackets around indices denote antisymmetrization and round brackets around indices denote symmetrization, for example, $\alpha_{[ab]} = \frac{1}{2}(\alpha_{ab} - \alpha_{ba})$, $\beta_{[abc]} = \frac{1}{6}(\beta_{abc} + \beta_{bca} + \beta_{cab} - \beta_{acb} - \beta_{cba} - \beta_{bac})$, $\gamma_{(AB)} = \frac{1}{2}(\gamma_{AB} + \gamma_{BA})$, etc.) Since (6.8), (6.10)–(6.12) involve German indices on the right-hand sides we have to show that the expressions defined are, in fact, basis-independent. This follows because the operators $\nabla_{[c}\nabla_{d]}$, $\Box_{AB'}$ $\Box_{A'B'}$ all satisfy

$$\begin{aligned} &\mathscr{D}(\xi^{\cdots}_{\cdots} + \eta^{\cdots}_{\cdots}) = \mathscr{D}\xi^{\cdots}_{\cdots} + \mathscr{D}\eta^{\cdots}_{\cdots} \\ &\mathscr{D}(\xi^{\cdots}_{\cdots}\zeta^{\cdots}_{\cdots}) = \zeta^{\cdots}_{\cdots}\mathscr{D}\xi^{\cdots}_{\cdots} + \xi^{\cdots}_{\cdots}\mathscr{D}\zeta^{\cdots}_{\cdots} \\ &\mathscr{D}\phi = 0 \end{aligned} \tag{6.13}$$

[*cf.* (6.16)], where $\mathscr{D}$ symbolically stands for any of these operators. To make a basis change in (6.8), (6.10)–(6.12) we make substitutions:

$$\delta^{\mathfrak{a}}_b \to T^{\mathfrak{a}}{}_{\mathfrak{x}}\,\delta^{\mathfrak{x}}_b \qquad \delta^a_{\mathfrak{a}} \to T^{(-1)\mathfrak{z}}{}_{\mathfrak{a}}\,\delta^a_{\mathfrak{z}} \tag{6.14}$$

$$\varepsilon_C{}^{\mathfrak{D}} \to t^{\mathfrak{D}}{}_{\mathfrak{X}}\,\varepsilon_C{}^{\mathfrak{X}} \qquad \varepsilon_{D\mathfrak{D}} \to t^{(-1)\mathfrak{Z}}{}_{\mathfrak{D}}\,\varepsilon_{D\mathfrak{Z}} \tag{6.15}$$

where (-1) denotes matrix inverse. By virtue of (6.13), the transforming matrix carries through the derivatives in each case and cancels with its inverse to leave the expression unchanged. (Note that abstract labels do not "transform"!)

The relation between these operators is given by

$$2\nabla_{[a}\nabla_{b]} \equiv \varepsilon_{A'B'}\Box_{AB} + \varepsilon_{AB}\Box_{A'B'} \tag{6.16}$$

so, for example, $\Box_{AB} \equiv \varepsilon^{A'B'}\ \nabla_{[a}\nabla_{b]}$. We can combine (6.10) and (6.12) in the form

$$\varepsilon_{D\mathfrak{D}}\Box_{AB}\varepsilon_C{}^{\mathfrak{D}} = \Psi_{ABCD} - 2\Lambda\varepsilon_{D(A}\varepsilon_{B)C} \tag{6.17}$$

Also, (6.17) combines with (6.11) to give (via (6.16))

$$2\varepsilon_{D\mathfrak{D}}\nabla_{[a}\nabla_{b]}\varepsilon_C{}^{\mathfrak{D}} = \Psi_{ABCD}\varepsilon_{A'B'} - 2\Lambda\varepsilon_{D(A}\varepsilon_{B)C}\varepsilon_{A'B'} + \Phi_{CDA'B'}\varepsilon_{AB} \tag{6.18}$$

In fact, Ψ_{ABCD}, Λ, and $\Phi_{ABC'D}$ are the parts [72, 115] of R_{abcd} *irreducible* under local $SL(2, \mathrm{C})$ (Lorentz) transformations:

$$\begin{aligned} R_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\Psi}_{A'B'C'D'} + 2\Lambda\{\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'B'}\varepsilon_{C'D'} \\ + \varepsilon_{AB}\varepsilon_{CD}\varepsilon_{A'D'}\varepsilon_{B'C'}\} + \varepsilon_{AB}\Phi_{CDA'B'}\varepsilon_{C'D'} + \varepsilon_{CD}\Phi_{ABC'D'}\varepsilon_{A'B'} \end{aligned} \tag{6.19}$$

Equation (6.19) follows by two applications of (6.18) [using (4.35)] to the definition (6.8). The symmetries

$$R_{abcd} = R_{cdab} = R_{[ab][cd]} \qquad R_{a[bcd]} = 0 \tag{6.20}$$

find expression in

$$\Psi_{ABCD} = \Psi_{(ABCD)} \qquad \Lambda = \bar{\Lambda} \qquad \Phi_{ABC'D'} = \Phi_{(AB)(C'D')} = \bar{\Phi}_{ABC'D'} \tag{6.21}$$

so that Ψ_{ABCD}, Λ, and $\Phi_{ABC'D'}$ belong, respectively, to a $D(2, 0)$, a $D(0, 0)$ and a $D(1, 1)$ irreducible representation space for the local Lorentz group.

In tensor terms, the reduction of R_{abcd} into irreducible parts is somewhat more complicated. We can define the *Weyl* (conformal) tensor by

$$C_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\Psi}_{A'B'C'D'} \tag{6.22}$$

Purely tensorially, this is given by

$$C^{ab}{}_{cd} = R^{ab}{}_{cd} - 2R^{[a}{}_{[c}\delta^{b]}_{d]} + \tfrac{1}{3}R\delta^{c}_{[a}\delta^{d}_{b]} \tag{6.23}$$

$$R_{ab} = R^{x}{}_{axb} \qquad R = R^{x}{}_{x} \tag{6.24}$$

Then, in addition to possessing all the Riemann tensor symmetries (6.20), C_{abcd} satisfies

$$C^{x}{}_{bxd} = 0 \tag{6.25}$$

The C_{abcd} is not yet irreducible (if we allow ourselves to consider complex tensors) and splits down further into its so-called self-dual and anti-self-dual parts: $C_{abcd} = C^{+}_{abcd} + C^{-}_{abcd}$, where $C^{+}_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'}$. ("Self-dual" means, here, $\frac{1}{2}C^{+}_{abxy}\,ie_{cd}{}^{xy} = C^{+}_{abcd}$, where $-e_{abc\mathfrak{d}}$ is the skew-symmetric Levi–Civita symbol in a right-handed orthonormal frame. In spinor terms $e_{abcd} = i\varepsilon_{AC}\varepsilon_{BD}\varepsilon_{A'D'}\varepsilon_{B'C'} - i\varepsilon_{AD}\varepsilon_{BC}\varepsilon_{A'C'}\varepsilon_{B'D'}$). The remaining irreducible parts of R_{abcd} are

$$R_{ab} - \tfrac{1}{4}Rg_{ab} = -2\Phi_{ABA'B'} \tag{6.26}$$

and

$$R = 24\Lambda \tag{6.27}$$

As a verification that the definition (6.8), of the Riemann tensor, actually agrees with the usual one, we can derive the *Ricci identities*. Consider, for example, $\xi^{a}{}_{b}$. Then,

$$\begin{aligned} 2\nabla_{[a}\nabla_{b]}\xi^{c}{}_{d} &= 2\nabla_{[a}\nabla_{b]}(\xi^{\mathfrak{c}}{}_{\mathfrak{b}}\delta^{\mathfrak{b}}_{d}\delta^{c}_{\mathfrak{c}}) \\ &= \xi^{c}{}_{x}R^{x}{}_{dab} - \xi^{x}{}_{d}R^{c}{}_{xab} \end{aligned} \tag{6.28}$$

by (6.13) and the fact [which also follows from (6.13) applied to (6.8)] that $R^{a}{}_{bcd} = -2\delta^{\mathfrak{a}}_{b}\nabla_{[c}\nabla_{d]}\delta^{a}_{\mathfrak{a}}$. In an exactly similar way we obtain the corresponding spinor Ricci identities, taking as an example

$$\begin{aligned} \Box_{A'B'}\theta_{C}{}^{D'} &= \Box_{A'B'}(\theta_{\mathfrak{C}}{}^{\mathfrak{D}'}\varepsilon_{\mathfrak{D}'}{}^{D'}\varepsilon_{C}{}^{\mathfrak{C}}) \\ &= \theta_{C}{}^{X'}(\bar{\Psi}^{D'}{}_{X'A'B'} - 2\Lambda\varepsilon_{X'(A'}\varepsilon_{B')}{}^{D'}) - \theta_{X}{}^{D'}\Phi^{X}{}_{CA'B'}. \end{aligned} \tag{6.29}$$

We can also directly obtain the expression for the components $R_{abc\mathfrak{d}}$ in terms of $\Gamma^{a}{}_{bc}$ by applying (6.1) to (6.8) [with (3.20)]. This includes both the familiar expression in terms of Christoffel symbols and the Ricci rotation coefficient curvature formula. Somewhat less familiar are the corresponding formulas which relate the spin coefficients to the components of the spinor curvature quantities. These will be given explicitly here, since we shall be interested in two particular cases later. Let us use (6.18) since this gives all the curvature quantities at once. The result is [64]:

$$\begin{aligned}\nabla_{\mathfrak{F}\mathfrak{E}'}\gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{D}\mathfrak{B}'} - \nabla_{\mathfrak{D}\mathfrak{B}'}\gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{F}\mathfrak{E}'} = \varepsilon^{\mathfrak{P}\mathfrak{Q}}\{&\gamma_{\mathfrak{A}\mathfrak{P}\mathfrak{D}\mathfrak{B}'}\gamma_{\mathfrak{Q}\mathfrak{C}\mathfrak{F}\mathfrak{E}'} + \gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{P}\mathfrak{B}'}\gamma_{\mathfrak{Q}\mathfrak{D}\mathfrak{F}\mathfrak{E}'} - \gamma_{\mathfrak{A}\mathfrak{P}\mathfrak{F}\mathfrak{E}'}\gamma_{\mathfrak{Q}\mathfrak{C}\mathfrak{D}\mathfrak{B}'}\\ &- \gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{P}\mathfrak{E}'}\gamma_{\mathfrak{Q}\mathfrak{F}\mathfrak{D}\mathfrak{B}'} + \varepsilon^{\mathfrak{R}'\mathfrak{S}'}\{\gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{D}\mathfrak{R}'}\gamma_{\mathfrak{S}'\mathfrak{B}'\mathfrak{E}'\mathfrak{F}} - \gamma_{\mathfrak{A}\mathfrak{C}\mathfrak{F}\mathfrak{R}'}\gamma_{\mathfrak{S}'\mathfrak{E}'\mathfrak{B}'\mathfrak{D}}\}\\ &+ \Psi_{\mathfrak{A}\mathfrak{C}\mathfrak{D}\mathfrak{F}}\varepsilon_{\mathfrak{E}'\mathfrak{B}'} + \Lambda\varepsilon_{\mathfrak{E}'\mathfrak{B}'}(\varepsilon_{\mathfrak{C}\mathfrak{D}}\varepsilon_{\mathfrak{A}\mathfrak{F}} + \varepsilon_{\mathfrak{A}\mathfrak{D}}\varepsilon_{\mathfrak{C}\mathfrak{F}}) + \Phi_{\mathfrak{A}\mathfrak{C}\mathfrak{B}'\mathfrak{E}'}\varepsilon_{\mathfrak{F}\mathfrak{D}} \qquad (6.30)\end{aligned}$$

In this connection, we note also the commutator of two "intrinsic derivatives":

$$\begin{aligned}\{\nabla_{\mathfrak{A}\mathfrak{B}'}\nabla_{\mathfrak{C}\mathfrak{D}'} - \nabla_{\mathfrak{C}\mathfrak{D}'}\nabla_{\mathfrak{A}\mathfrak{B}'}\}\phi = \{&\varepsilon^{\mathfrak{P}\mathfrak{Q}}(\gamma_{\mathfrak{P}\mathfrak{A}\mathfrak{C}\mathfrak{D}'}\nabla_{\mathfrak{Q}\mathfrak{B}'} - \gamma_{\mathfrak{P}\mathfrak{C}\mathfrak{A}\mathfrak{B}'}\nabla_{\mathfrak{Q}\mathfrak{D}'})\\ &+ \varepsilon^{\mathfrak{R}'\mathfrak{S}'}(\gamma_{\mathfrak{R}'\mathfrak{B}'\mathfrak{D}'\mathfrak{C}}\nabla_{\mathfrak{A}\mathfrak{S}'} - \gamma_{\mathfrak{R}'\mathfrak{D}'\mathfrak{B}'\mathfrak{A}}\nabla_{\mathfrak{C}\mathfrak{S}'})\phi \qquad (6.31)\end{aligned}$$

7 EINSTEIN'S EQUATIONS AND FOCUSING

The decomposition (6.19) of the Riemann tensor into its irreducible spinorial parts allows us to discuss the structure of space-time curvature, as is implied by Einstein's field equations, in some detail. Einstein's equations are

$$R_{ab} - \tfrac{1}{2}Rg_{ab} + \lambda g_{ab} = -2KT_{ab} \qquad (7.1)$$

where λ and K are real constants with $K > 0$ and where T_{ab} is the local stress–energy–momentum tensor (the "energy tensor") of matter. Here, the term "matter" refers to any field except gravitation itself. Thus, for example, free electromagnetic field counts as "matter" in this context. The question of the energy which must be assigned to gravitation *itself* is a more subtle one. This will be discussed a little more later. Gravitational energy cannot be adequately defined in a *local* way and emerges, instead, as some kind of non-local quantity. The local gravitational energy must apparently be thought of as zero, and this is consistent with (7.1). The cosmological term λ is generally put equal to zero, since there are no compelling theoretical or observational grounds for believing in its existence. In any case, λ would have to be extremely small, of the order of the reciprocal of the square of the radius of the universe.

As it stands, (7.1) implies no restriction on the curvature, since no restriction has yet been placed on T_{ab}. But we can study (7.1) in three different types of situation. In the first instance, we may be interested in *pure gravitation theory* for which $T_{ab} = 0$. Then (7.1) becomes a well-defined set of partial differential nonlinear equations on the g_{ab}, when expressed in terms of a coordinate system x^a. These equations are difficult to solve explicitly and only a very few solutions are known which have any direct relevance to physical situations. (Some of these will be given in later sections.) On the other hand, there are certain general statements which can be made about the nature (in particular the asymptotic nature, if suitable boundary conditions are imposed) of the solutions. Now, secondly, we may be interested in solutions of (7.1) for which the T_{ab}, though nonzero, is also subject to certain conditions ("equations of state") which are normally again partial differential equations, but which now refer to the field quantities which contribute to T_{ab}. (For example, in the case of the "Einstein-Maxwell" equations, T_{ab} is specified as being the energy tensor of a free electromagnetic field which is subject to Maxwell's source-free equations in covariant form.) Again the equations become virtually impossible to solve explicitly, in general, for realistic equations of state. Thirdly, we may be content to make general statements about solutions of (7.1) for which we merely impose *inequalities* on the T_{ab}. Such an inequality might state, for example, the positive definite nature of the local energy density, or some other such physically reasonable requirement. In many ways, this type of situation is the easiest to handle and the last two sections will be devoted mainly to the treatment of space-times under such conditions. Finally, it may be remarked that one way of attempting to treat (7.1) which strictly speaking is *not* valid, is to regard T_{ab} as some *given* source distribution for which we try to construct its gravitational field. This is a normal procedure in other branches of physics, but here, *until* we have our metric g_{ab} it is not even clear what is *meant* by a given source distribution.

If we now decompose R_{abcd} into its spinorial parts and use (6.26) and (6.27), we get

$$KT_{ab} = \Phi_{ABA'B'} + (3\Lambda - \tfrac{1}{2}\lambda)\varepsilon_{AB}\varepsilon_{A'B'} \tag{7.2}$$

so Φ_{ab} represents the trace-free part of T_{ab} and Λ the trace:

$$\Phi_{ab} = K(T_{ab} - \tfrac{1}{4}T_x{}^x g_{ab}) \qquad \Lambda = \tfrac{1}{12}(KT_x{}^x + 2\lambda) \tag{7.3}$$

The part of R_{abcd} which, locally, is left completely *undefined* by T_{ab} is the Weyl tensor C_{abcd}, or, equivalently, the spinor Ψ_{ABCD}. We may thus think of Ψ_{ABCD} as representing the *free gravitational part of the curvature*. In regions free of matter we can still have curvature. According to Einstein's equations, then, such curvature will be of a particular type, defined completely, in structure, by a totally symmetric spinor Ψ_{ABCD}. If $\lambda = 0$, then we shall have, in fact, $R_{abcd} = C_{abcd}$ [with C_{abcd} given by (6.22)]. Let us, therefore, examine the nature of Ψ_{ABCD}. Spinor methods are particularly well suited to this problem, the corresponding discussion of the Weyl tensor C_{abcd} being far less transparent.

Consider the form $\Psi_{\mathfrak{ABCD}}\xi^{\mathfrak{A}}\xi^{\mathfrak{B}}\xi^{\mathfrak{C}}\xi^{\mathfrak{D}}$, where for simplicity we choose $\xi^0 = 1$, $\xi^1 = z$. This is a quartic polynomial in z with complex coefficients and therefore factorizes into four linear factors:

$$\Psi_{\mathfrak{ABCD}}\xi^{\mathfrak{A}}\xi^{\mathfrak{B}}\xi^{\mathfrak{C}}\xi^{\mathfrak{D}} = (\alpha_{\mathfrak{A}}\xi^{\mathfrak{A}})(\beta_{\mathfrak{B}}\xi^{\mathfrak{B}})(\gamma_{\mathfrak{C}}\xi^{\mathfrak{C}})(\delta_{\mathfrak{D}}\xi^{\mathfrak{D}}) \tag{7.4}$$

Equating coefficients [and eliminating the basis frame *cf.* (3.21)] we get

$$\Psi_{ABCD} = \alpha_{(A}\beta_B\gamma_C\delta_{D)} \tag{7.5}$$

The factorization is unique up to a complex multiplier for each of $\alpha_A, \ldots, \delta_D$. Thus, the null direction defined by each of $\alpha_A, \ldots, \delta_D$ is uniquely defined, so we see that Ψ_{ABCD} defines an unordered set of four (possibly coincident) directions along the null cone at each point P of M (at which $\Psi_{ABCD} \neq 0$) called the *gravitational principal null directions* [23, 72, 91] at P. The coincidence schemes for the principal null directions at P can be represented as follows:

$$\begin{tikzcd}
\{1111\} \arrow[r] & \{211\} \arrow[dl] \arrow[dr] & \\
\{22\} \arrow[dr] & & \{31\} \arrow[dl] \\
& \{4\} \arrow[dl] & \\
\{-\} & &
\end{tikzcd} \qquad (7.6)$$

Petrov type: I II III

where the arrows point in the direction of increasing specialization. Here, $\{1111\}$ represents the general case when the null directions are distinct; $\{211\}$ is the case when there is one double null direction and two distinct simple ones; etc. The symbol $\{-\}$ denotes the case $\Psi_{ABCD} = 0$ at P. The "Petrov type," referring to the columns in (7.6), is a more direct classification [80b, 81] of C_{abcd} according to the dimension of the space spanned by eigenvectors of the matrix $(C^{\mathfrak{ab}}{}_{\mathfrak{cd}})$ (grouping $\mathfrak{a}, \mathfrak{b}$ together and $\mathfrak{c}, \mathfrak{d}$ together). The case $\{4\}$, when all the principal null directions coincide, is called the *null* case; $\{22\}$ is called *degenerate*; all cases but $\{1111\}$ are called *algebraically special*. In terms of C_{abcd}, we can write a necessary and sufficient condition [23,93] on a null vector l^a for it to point in a principal null direction as

$$l_{[a}C_{b]cd[e}l_{f]}l^c l^d = 0 \tag{7.7}$$

The classification scheme (7.6) plays a significant part in the understanding of the geometric structure of gravitational fields. Perhaps the most directly physical role of the principal null directions emerges in the "peeling-off" property of Sachs [64, 76, 93–95]. This states that for an asymptotically flat space-time which is empty near infinity (definitions in Section 8), then along any null geodesic γ, the curvature tensor exhibits a certain characteristic asymptotic behavior. Let r be an affine parameter on γ. Then, in a well-defined sense, the curvature falls off along γ in such a way that to order r^{-1} it is null, with the quadruple principle null direction pointing along γ; to order r^{-2}, there is a triple principal null direction along γ; to order r^{-3}, a double one; to order r^{-4}, a simple one; and to order r^{-5} the curvature is unrelated to γ (Fig. 9). This is the generic situation and more special behavior is also possible. An outline of a proof of this result is given in Section 8.

CHAPTER 36

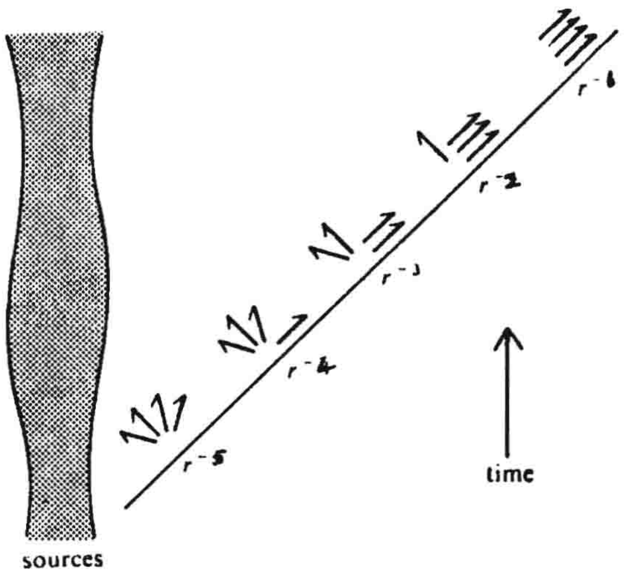


FIGURE 9. The peeling-off property.

The decomposition (6.19) was given purely algebraically. We may ask whether there is a more direct geometrical (or "physical") way of dinstinguishing the effect of the Weyl tensor from that of other parts of the curvature. Also, can one see physically what the meaning of the principal null directions is? The most direct physical manifestation of space-time curvature is to be found in the "tidal forces" due to the geodesic deviation effect on timelike geodesics [82, p. 266]. The relevant equation can be written

$$D^2 p^a - R^a{}_{bcd} t^b p^c t^d = 0 \tag{7.8}$$

where the t^a are tangents to a congruence of timelike geodesics, smoothly parametrized according to proper time s, say, so $t^a t_a = 1$. The vector p^a connects points of equal s value on two neighboring geodesics (Fig. 10). The operator D denotes covariant differentiation in the direction t^a, that is,

$$D \equiv t^a \nabla_a \tag{7.9}$$

The equation (7.8) is familiar to mathematicians as the *Jacobi equation.* [Actually (7.8) holds under more general conditions than those stated, namely, that the parameter s need only be affine on each geodesic.] If we choose p^a orthogonal to t^a (as we may) then $D^2 p^a$ measures the *relative acceleration* of neighboring intertial test particles. [A simple "intuitive derivation" of (7.8), when the two neighboring geodesics are initially parallel, is obtained by simply transporting the vector t^a around the loop $RQPSR$ of Fig. 10. The change in t^a gives us the change in the relative velocity of the neighboring particle.]

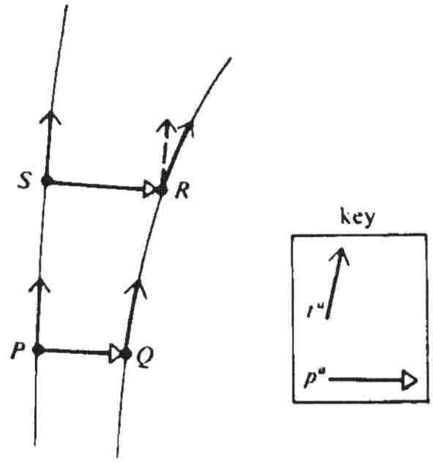


FIGURE 10. The relative acceleration of neighboring geodesics in the presence of curvature.

As applied to timelike geodesics, (7.8) does not enable us to separate, in any simple way, the effects due to the different irreducible parts of the curvature tensor. But if we use null geodesics, the effects due to the Weyl tensor and due to the (trace-free) Ricci tensor become sharply distinguished [83,102]. Since the case of null geodesics has some peculiar features which are rather suitable to a spinor treatment, I shall give an independent spinor derivation of the relevant equations, rather than trying to deduce them from (7.8). This will also give us an interpretation of some of the spin coefficients, which will be useful later.

Choose a spin frame $\varepsilon_{\mathfrak{A}}{}^{A}$ and put

$$\varepsilon_0{}^A = o^A \qquad \varepsilon_1{}^A = \iota^A \tag{7.10}$$

From the normalization (4.27) we get

$$o_A \iota^A = 1 \tag{7.11}$$

Let us try to interpret the following spin coefficients [*cf.* (6.2)]:

$$\kappa = \gamma_{000 0'} \qquad \rho = \gamma_{001 0'} \qquad \sigma = \gamma_{000 1'} \tag{7.12}$$

There are all of the form $o^A \nabla_{\mathfrak{C}\mathfrak{D}'} o_A$, where $\mathfrak{C}$, $\mathfrak{D}'$ are, respectively, 0, 0′; 1, 0′; 01′. Thus, under

$$o^A \to \zeta o^A \qquad \iota^A \to \zeta^{-1}\iota^A \tag{7.13}$$

we have

$$\kappa \to \zeta^3\bar{\zeta}\kappa \qquad \rho \to \zeta\zeta\rho \qquad \sigma \to \zeta^3\bar{\zeta}^{-1} \tag{7.14}$$

the derivative terms canceling because of $o^A o_A = 0$. Now, consider the meaning of the condition $\kappa = 0$:

$$o^A o^B \bar{o}^{C'} \nabla_{BC'} o_A = 0 \tag{7.15}$$

We can read this as something of the form $o^A 0_A = 0$, from which we can infer $0_A = \chi o_A$, for some χ. Thus $\kappa = 0$ means

$$o^B\bar{o}^{C'}\nabla_{BC'}o_A = \chi o_A \tag{7.16}$$

Putting

$$l^a = o^A\bar{o}^{A'} \qquad m^a = o^A\bar{\iota}^{A'} \tag{7.17}$$

we rewrite (7.16) as $l^a\nabla_a o_A = \chi o_A$, which states that o_A is carried parallel to itself in the l^a-direction (that is, in the o^A-direction). That is to say, the condition $\kappa = 0$ is necessary and sufficient for the o^A-directions to be tangent to *null geodesics*. Thus, κ is a kind of measure of *curvature* of the o^A-curves. (It is complex, so it also indicates the *direction* of this curvature, in relation to the flag plane of o^A.)

Suppose now that $\kappa = 0$, so we have a congruence of null geodesic o^A-curves, and read (7.15) in two other ways, namely, as $o^B\phi_B = 0$ or as $\bar{o}^{C'}\psi_{C'} = 0$. Thus, similarly to the above, we have relations of the form

$$o^A\bar{o}^{C'}\nabla_{BC'}o_A = \rho o_B \tag{7.18}$$

$$o^A o^B\nabla_{BC'}o_A = \sigma\bar{o}_{C'} \tag{7.19}$$

for some ρ,σ. In fact, ρ and σ are the quantities given in (7.12) as follows at once by multiplying (7.18) by ι^B and (7.19) by $\bar{\iota}^{C'}$. The significance of (7.18) and (7.19) is that it shows that when $\kappa = 0$, the quantities ρ and σ are independent of ι^A. Also, since they are simply rescaled according to (7.14), as o^A undergoes (7.13), we expect to find a simple geometric meaning for ρ and σ, as a property of the congruence of null geodesics. To see that this is actually the case, consider a 2-plane element π at a point P of one of the geodesics γ. Choose π spacelike and orthogonal to the direction of γ. We can select ι^A also to have direction orthogonal to π, so π will be spanned by m^a and $\bar{m}^a$, with m^a as in (7.17). Now any (real) vector in π has the form

$$\bar{z}m^a + z\bar{m}^a \tag{7.20}$$

so we may regard π as being the Argand plane of $\sqrt{2}\,z$. In the neighborhood of γ, a two-dimensional set of the null geodesics "hit" π. These are the null geodesics which lie, to first order, in a null hyperplane through γ. There are also other null geodesics which are neighboring to γ, but we do not consider them. We shall examine how z varies as we follow one of the null geodesics which does hit π. [Note that the z, at P, for a particular null geodesic of the congruence, near γ, is actually independent of the choice of π at P. Altering π through P amounts merely to replacing m^a by $m^a + kl^a$ in (7.20), with k complex.]

Put

$$D \equiv l^a\nabla_a \equiv \nabla_{00'} \tag{7.21}$$

For convenience, choose o^A, ι^A to be parallelly propagated along γ:

$$Do_A = 0 \qquad D\iota_A = 0 \tag{7.22}$$

Then m^a, $\bar{m}^a$, and therefore also π, will be parallelly propagated along γ. Let ϕ be any function constant along the null geodesics:

$$D\phi = 0 \tag{7.23}$$

We characterize the fact that we require a fixed value of z to label a definite geodesic of our system neighboring γ by

$$D\{(\bar{z}m^a + z\bar{m}^a)\nabla_a\phi\} \equiv 0 \tag{7.24}$$

[see (7.20)]. This gives

$$D\bar{z}\nabla_{01'}\phi + Dz\nabla_{10'}\phi + \bar{z}(\nabla_{00'}\nabla_{01'} - \nabla_{01'}\nabla_{00'})\phi + z(\nabla_{00'}\nabla_{10'} - \nabla_{10'}\nabla_{00'})\phi = 0 \tag{7.25}$$

that is, by (6.31) (or directly evaluating these particular commutators)

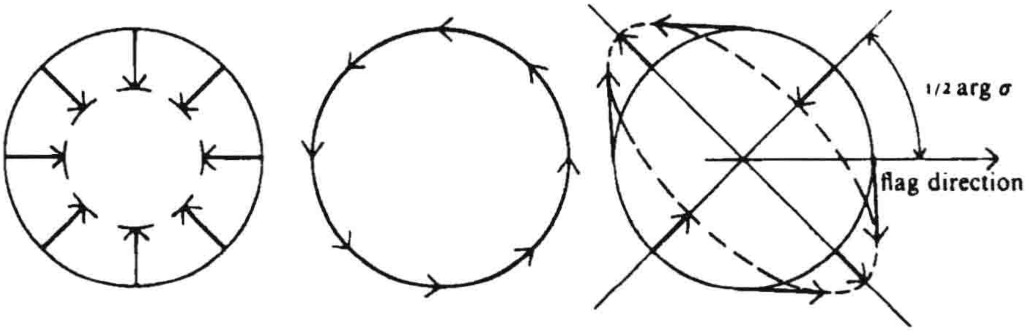
$$(D\bar{z} + \bar{\rho}\bar{z} + \bar{\sigma}z)\nabla_{01'}\phi + (Dz + \rho z + \sigma\bar{z})\nabla_{01'}\phi = 0 \tag{7.26}$$

Thus

$$Dz = -\rho z - \sigma\bar{z} \tag{7.27}$$

The interpretation of ρ and σ now becomes clear. The *real* part of ρ measures the *convergence* of the null geodesics and the *imaginary* part of ρ measures their *rotation* about γ. A nonvanishing σ indicates the presence of distortion or *shear* (Fig. 11). Thus, if $\sigma \neq 0$, a small circle will become elliptical as we follow the null geodesics. The axes of this ellipse are defined in relation to the flag plane of o^A by $\frac{1}{2}\arg\sigma$. If the null geodesics are the generators of null hypersurfaces then we can choose $l_a = \nabla_a u$, whence

$$o^A\bar{o}^{B'}(\nabla_a l_b - \nabla_b l_a) = 0$$



Re ρ Im ρ σ

FIGURE 11. The interpretation of ρ and σ in terms of their effect on an initially circular pattern in the null rays.

from which we get absence of rotation:

$$\rho = \bar{\rho} \tag{7.28}$$

Condition (7.28) is also sufficient for the null geodesics to generate null hypersurfaces. If (7.28) holds, then the ρ and σ pertain to the geometry of each null hypersurface individually and not to the relationship among the different null hypersurfaces.

Let us next consider the D-derivatives of ρ and σ. We can obtain these from (6.30) [using (7.22): $\gamma_{\mathfrak{a}\mathfrak{b}00'} = 0$], or else directly, using the definitions of ρ, σ, $\Phi_{ABC'D'}$, and Ψ_{ABCD}. Thus,

$$D\rho = \rho^2 + \sigma\bar{\sigma} + \Phi \tag{7.29}$$

$$D\sigma = \sigma(\rho + \bar{\rho}) + \Psi \tag{7.30}$$

where

$$\Psi = \Psi_{0000} \qquad \Phi = \Phi_{000'0'} \tag{7.31}$$

Equations (7.29) and (7.30) are the *Sachs equations* [93] for the "optical scalars" ρ, σ. Combining these with (7.27) we get

$$D^2 z = -\Phi z - \Psi \bar{z} \tag{7.32}$$

In fact, (7.32) can also be obtained directly from (7.8), if we put $t^a = l^a$ and $p^a = \bar{z}m^a + z\bar{m}^a$. In this connection, we note the formulas

$$\begin{aligned} \Psi &= R_{abcd}\, l^a m^b l^c m^d & \Phi &= R_{abcd}\, l^a m^b l^c \bar{m}^d; \\ \rho &= m^a \bar{m}^b \nabla_b l_a & \sigma &= m^a m^b \nabla_b l_a \end{aligned} \tag{7.33}$$

We see from (7.29) that it is essentially Φ which governs the change in ρ and from (7.30) that it is essentially Ψ which governs the change in σ, but that in both cases there are additional nonlinear terms. To obtain the effects of Φ and Ψ in purest form, consider a particular situation in which Φ and Ψ are localized to a very small region on γ. [In fact idealized situations are also possible for which Φ and Ψ become Dirac delta functions. See the discussion of plane waves (9.23)–(9.26).] Suppose that we have a pulse of light which is initially a parallel beam. This is represented by the null geodesics constituting a small portion (neighboring γ) of a null hypersurface through γ, for which $\rho = \sigma = 0$. Then, by (7.29), we see that a patch of Φ curvature (with $\Psi = 0$) acts as a *lens without astigmatism* (since convergence ρ is introduced into the beam but the shear σ remains zero) while a patch of Ψ curvature (with $\Phi = 0$) acts as a *purely astigmatic lens*; that is, exactly as much positive convergence is introduced in one plane as negative convergence in the perpendicular plane (since ρ is initially zero for the emergent beam with σ nonzero). The interpretation of a patch of curvature in terms of lenses also holds good for a general incident beam, where ρ and σ needs not initially vanish. Also, the nonlinear terms in (7.29) and (7.30) correspond exactly to the way that the effects of separated lenses compose [78]. Thus, we may think of the effect on a beam of light, of the curvature of space-time, as being built up precisely from effects due to a series of lenses placed along the beam. The converging power is described locally by Φ and the astigmatism by Ψ. As γ varies, the components (7.31) generate *all* the information contained in Ψ_{ABCD} and $\Phi_{ABC'D'}$, because of the symmetries (6.21).

One important feature of this is that we always have *positive* focusing along any null geodesic γ, if we assume that the local energy density is positive definite. The energy density, measured by an observer whose world line has tangent vector t^a, is

$$T_{ab}t^at^b \tag{7.34}$$

if we assume $t^at_a = 1$. If we require (7.34) to be nonnegative, then we have $T_{ab}t^at^b \geqslant 0$ for any timelike vector t^a (irrespective of the normalization condition on t^a). Thus, if we take the limit as t^a approaches a null vector l^a, we get [by (7.3)]

$$\Phi = KT_{ab}l^al^b \geqslant 0 \tag{7.35}$$

(whether or not $\lambda = 0$, since $g_{ab}l^al^b = 0$). In the absence of Ψ, this shows that the curvature Φ along the beam acts as a series of positive (converging) lenses. If Ψ is present, its effect is made felt in Eq. (7.29) via the $\sigma\bar{\sigma}$ term, which again acts positively. We can regard σ as being, roughly speaking, an *integral* of Ψ along γ. Thus, Ψ acts *nonlocally* to contribute to the focussing power along γ. An oscillating Ψ closely resembles the presence of Φ in its total effect. This is closely related to the question of gravitational energy. Although locally the gravitational field does not contribute to the energy density, a gravitational wave does carry a kind of nonlocal energy, the total energy carried by the wave being always *positive*. (A striking example of the effective positive energy of gravitational "ripples" is supplied by Wheeler's [112] theoretical construct: the gravitational "geon." This is a region of empty space which consists of ripples which remain bound for a long time. When viewed on a large enough scale, a geon appears as a "particle" with positive mass.)

Let us examine the positive focusing effect in a little more detail. If we choose a small triangle in the plane element π with vertices defined by 0, z_1, z_2, then its area

$$a = \frac{i}{4}(z_1\bar{z}_2 - z_2\bar{z}_1) \tag{7.36}$$

satisfies [by (7.27)]

$$Da = -(\rho + \bar{\rho})a \tag{7.37}$$

With our beam chosen as a portion of a null hypersurface we have $\rho = \bar{\rho}$. Thus,

$$D(a^{1/2}) = -\rho a^{1/2} \tag{7.38}$$

From (7.29),

$$D^2(a^{1/2}) = D(-\rho a^{1/2}) = -(\sigma\bar{\sigma} + \Phi)a^{1/2} \leqslant 0 \tag{7.39}$$

Thus, if $\rho > 0$ at some point of γ, (7.38) tells us that $a^{1/2}$ is decreasing. From (7.39) it follows that $a^{1/2}$ decreases to zero[20] (see Fig. 12). Thus the beam *inevitably reaches a focal point* Q. (This will in general be an astigmatic focal point where the beam collapses locally to a line rather than a point.) Note that, near Q, ρ *becomes unbounded* since by (7.38), $\rho = -1/2D\,(\log a)$.

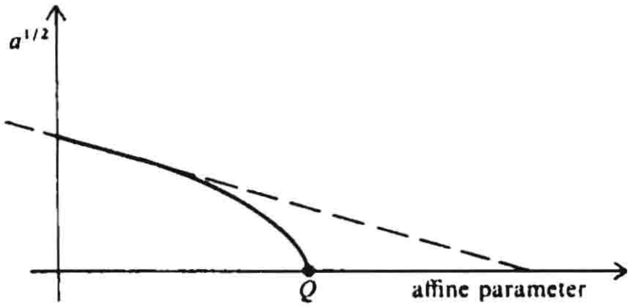


FIGURE 12. Once the local area of cross section of the null rays begins to decrease, it inevitably decreases to zero.

This behavior is a limiting case of the *Raychaudhuri effect* [49, 86], which applies to a hypersurface orthogonal timelike congruence of geodesics. Let t^a be the tangent vectors to the geodesics of the congruence, parametrized according to proper time s (so $t^a t_a = 1$; with t^a future-pointing) and orthogonal to a spacelike hypersurface $s = 0$. Define the divergence of the geodesics by

$$\theta = \nabla_a t^a \tag{7.40}$$

From the Ricci identities we obtain Raychaudhuri's equation

$$D\theta = -\nabla_a t^b \nabla_b t^a + R_{ab} t^a t^b \tag{7.41}$$

with $D = t^a \nabla_a$ as in (7.9).

---

[20] There is one simple application of this property which does not seem to have been noted before, namely, to the question of whether a spherically symmetrical body in asymptotically flat space-time can fail to have a center (see [9] p. 436). If we consider the light rays (null goeodesics) converging inward from an instantaneous sphere situated symmetrically about the body at a reasonable distance from it, we see that the rays initially converge, so that $\rho > 0$. They have to reach a (symmetrical) focal point (assuming (7.35) holds) and this focal point defines the center. (Compare the discussion of trapped surfaces given in Section 10.)

From the hypersurface orthogonality we get $\nabla_a t_b = \nabla_b t_a$, whence

$$\nabla_a t^b \nabla_b t^a = \nabla_a t_b \nabla^a t^b \geqslant \tfrac{1}{3}\theta^2 \tag{7.42}$$

from Schwarz' inequality in the 3-space orthogonal to t^a [by the geodesic condition $t^a \nabla_a t_b = 0$ and $t^a \nabla_b t_a = \frac{1}{2}\nabla_b(t^a t_a) = 0$]. Thus,

$$D\theta \leqslant -\tfrac{1}{3}\theta^2 + R_{ab}t^a t^b \tag{7.43}$$

Let V be a small 3-volume element orthogonal to t^a. Then

$$DV = \theta V \tag{7.44}$$

by (7.40), so

$$D(V^{1/3}) = \tfrac{1}{3}\theta V^{1/3} \tag{7.45}$$

and (7.43) gives

$$D^2(V^{1/3}) = \tfrac{1}{3}D(\theta V^{1/3}) \leqslant \tfrac{1}{3}V^{1/3}R_{ab}t^a t^b \leqslant 0 \tag{7.46}$$

if we assume that the *energy condition*

$$T_{ab}t^a t^b \geqslant \frac{\lambda}{2K} + T^c{}_c \tag{7.47}$$

holds [*cf.* (7.1)]. (If $\lambda = 0$, then (7.47) holds if in an "eigentetrad" of T_{ab}, the energy density is not less than *minus* each principal pressure, nor less than minus the sum of the principal pressures. This would be true of all normal matter.) At this point the argument proceeds just as before, with (7.45) replacing (7.38) and (7.46) replacing (7.39), showing that if $\theta < 0$ at some point on one of the geodesics, then a focal point ($V = 0$, $\theta \to -\infty$) is reached somewhere in the future on the geodesic. Similarly, if $\theta > 0$, then a focal point ($V = 0$, $\theta \to +\infty$) is reached in the past on the geodesic.

One final remark concerning (7.30) is appropriate here. We have, by (7.31),

$$\Psi = \Psi_{ABCD}\, o^A o^B o^C o^D \tag{7.48}$$

so, comparing with (7.4) and (7.5) we see that Ψ vanishes when o^A points in a *principal null direction*. Thus, we can give a physical interpretation of the principal null directions as those null directions (at a point P) along which *no astigmatic focusing* takes place. From the fact that astigmatism is a directional quantity, we can actually see from topological considerations that there must be at least a net algebraic count of *four* null directions of vanishing astigmatism at P. We consider an S^2 representing the null directions at P, on which we put a small line element to represent the astigmatism for each null direction at P. The principal null directions emerge as four points where the line element vanishes, as indicated in Fig. 13. This type of argument applies to other spins (for example, to spin 1, namely, electromagnetism, when the analogous line element is *oriented*, so the Fig. 13 behavior cannot

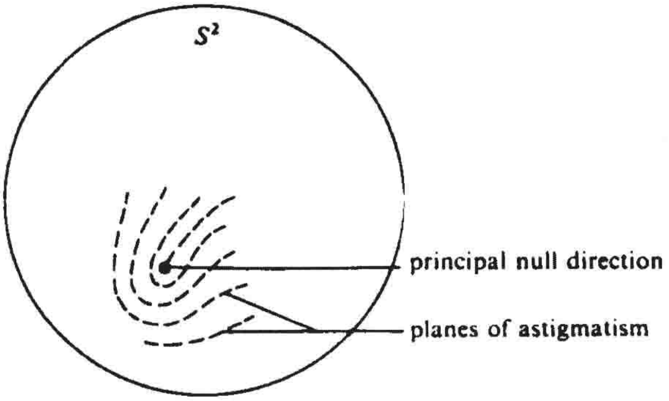


FIGURE 13. A principal null direction appears as a singularity in the field of directions of astigmatism, on the S^2 representing null directions at P.

occur and we get just two principal null directions). The argument can also be used in other situations, such as for the *asymptotic* fields for which S^2 would be at infinity rather than at P.

8 CONFORMAL INFINITY

The question of the meaning of gravitational energy has been briefly touched upon in the previous section. The utility of the concept of energy, in general, arises from the fact that it is *conserved*. In Einstein's theory there is a "local conservation law" for energy built in, namely,

$$\nabla_a T^{ab} = 0 \tag{8.1}$$

this being, by virtue of the field equations (7.1), a double contraction of the Bianchi identities:

$$\nabla_{[a} R_{bc]de} = 0 \tag{8.2}$$

However, (8.1) does *not* yield an *integral* conservation law of the usual type, that is, a law stating that an integral over the boundary of some compact 4-volume of the "flux" of some quantity across this boundary necessarily vanishes.

The prototype of such an integral conservation law is given by electric charge. If J^a denotes the charge current vector, then we have

$$\nabla_a J^a = 0 \tag{8.3}$$

Since J^a is a "vector" or, more appropriately, the "dual of a 3-form," the local conservation law (8.3) implies the existence of an integral conservation law of the above type, where the flux of charge across a surface element

dS_a is $J^a dS_a$. However, (8.1) is not of this form. We can get an *approximate* integral conservation law if we integrate over the boundary of some region whose dimension is very small compared with the radii of curvature involved in R_{abcd}. For this, we may introduce a basis frame $\delta^a_{\mathbf{a}}$ which is approximately covariantly constant over the region, and read (8.1) as $\nabla_{\mathbf{a}} T^{\mathbf{a}b} \bumpeq 0$. This is now of the form (8.3), so an approximate energy-momentum integral conservation law arises with energy-momentum flux $T^{ab} dS_a$. We may regard the space-time curvature as giving a (nonlocal) *gravitational* contribution to the energy-momentum which has to be taken into account to get an exact integral conservation law. This can be formalized in various ways (by the use of "pseudotensors", etc.), but no method has yet emerged which, in general situations, assigns any kind of meaningful uniqueness to the gravitational energy-momentum flux, say, or to the total energy momentum in a region. I think we may regard the question of gravitational energy as one of the important unsolved problems in general relativity theory.

But have we any right to expect a solution to this problem at all? Might it not be that energy-momentum is simply not quite conserved in general relativity, or, put another way, that the energy-momentum concept is not (except locally) really an appropriate one for the subject. It may be that in general this is so. I do not know. However, it is (to my mind) a very remarkable fact that a meaningful (that is, "covariantly" or "geometrically" defined) energy-momentum concept really *does* exist for an interesting subclass of space-times, namely, those which are *asymptotically flat*.

The study of asymptotically flat space-times forms an important part of general relativity theory. This is not because it is believed that the universe is very likely to be asymptotically flat, but because in any situation in which general-relativistic effects are expected to be important—that is, with the notable exception of the study of the universe as a whole—the curvatures involved in the local process will exceed by many orders of magnitude the smoothed-out curvatures of the general background. (For example, even if we collapse an entire galaxy, the size of the resulting object of interest would be smaller than the radius of the background curvature of the universe by a factor of about 10^{11}. See the discussion in Section 10.) Thus, asymptotic flatness is an excellent approximation in a large number of situations. This is fortunate, because it is only with asymptotic flatness that general relativity begins to resemble much of the rest of physics. We can then discuss the question of advanced and retarded waves, gravitational or otherwise. We can discuss the energy carried by these waves, scattering problems, even perhaps quantization. I shall not go into any of this in any detail here. I merely describe a mathematically precise framework for the discussion of these questions and just indicate some of the applications.

In order to fix ideas, let us consider, in more physical terms, the question of whether gravitational waves emitted by an isolated system carry energy, and if they do, whether the energy they carry is necessarily positive. (The existence of gravitational waves emitted by a system of varying asymmetry, is inferred in the first instance, from the linear approximation to Einstein's theory.) To find out what energy we should assign to the waves, we envisage measuring the mass m_1 of the system before the waves are emitted and then measuring the mass m_2 of the system after the waves are emitted. If energy (that is, mass) conservation is to retain any meaning, we require that the total energy carried by the waves be just $m_1 - m_2$ (assuming incoming waves are not present, which might be absorbed by the system).

But how are we to measure the mass? One way might be to integrate some expression of mass density over a spacelike hypersurface $\mathscr{S}$. However, we would have to take into account the "nonlocal mass density" of the gravitational field itself and this complicates things. As an alternative, we might prefer to measure the mass by examining the field *only* in the neighborhood of *infinity* on $\mathscr{S}$, since the way in which the curvature falls off at large distances on $\mathscr{S}$ should fix the total mass. Instead of a 3-volume integral, we then use just a 2-surface integral at infinity. In fact, either method *can* be used to give a satisfactory definition. But in both cases, our spacelike hypersurface $\mathscr{S}$ has to extend to infinity and we have a difficulty when it comes to measuring the mass $m_1 - m_2$ carried by the waves. We would "normally" require that $\mathscr{S}$ should suitably tend to become a spacelike hyperplane as our space-time approaches flatness at infinity. This is not unreasonable for the hypersurface $\mathscr{S}_1$, which intercepts the source's world line before the emission of the radiation, to yield m_1 (Fig. 14). But in the case of such a hypersurface $\mathscr{S}_2$, which intercepts the source *after* the emission of the radiation, we do not get m_2 as our mass measurement, but simply m_1 again. This is because $\mathscr{S}_2$ intercepts all the waves, and *whatever* energy they carry has to be added in, giving $m_1 = m_2 + (m_1 - m_2)$.

We might try to carry out some complicated limiting process, allowing the hypersurface to move "upward" into the future while letting the size of the region of integration extend gradually out to infinity, the waves remaining outside the region. Let us apply this to the second of the above two suggestions for measuring the mass. Then we are led to considerations of 2-surface integrals at infinity, not in spacelike directions, but in *null directions*. This suggests the use of null (or, at least, asymptotically null) hypersurfaces $\mathscr{N}_1$, $\mathscr{N}_2$, in place of $\mathscr{S}_1$, $\mathscr{S}_2$. Here $\mathscr{N}_1$ lies entirely outside the cone of radiation, which in turn lies entirely outside $\mathscr{N}_2$. Then, a 2-surface integral at infinity on $\mathscr{N}_i$ might reasonably be used to measure $m_i (i = 1, 2)$, the difference $m_1 - m_2$ then giving the mass of the waves.

By achieving such a definition of mass and by showing that the difference $m_1 - m_2$ is positive in the presence of radiation, Bondi (with his co-workers) and subsequently, Sachs [8, 11, 95], made an important advance in

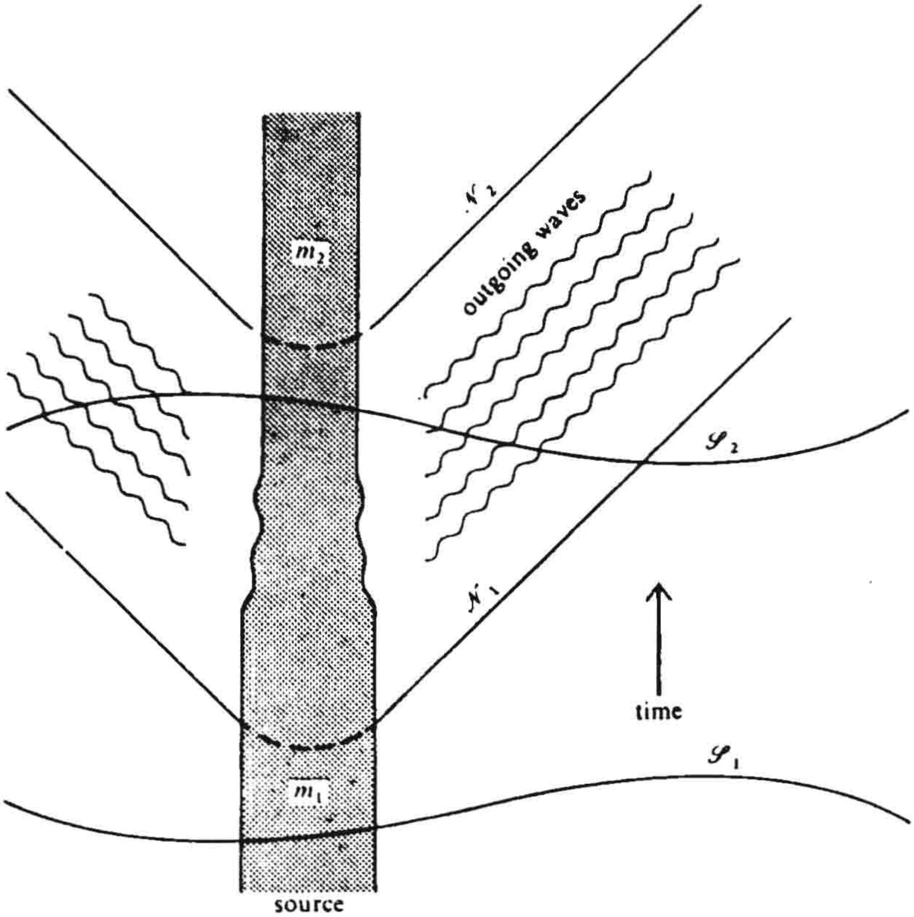


FIGURE 14. To measure mass loss through radiation, $\mathcal{N}_1$ and $\mathcal{N}_2$ are more appropriate than $\mathcal{S}_1$ and $\mathcal{S}_2$.

the understanding of gravitational energy. Their definition depended on the introduction of a specific type of coordinate system based on outgoing null hypersurfaces. The existence of a coordinate system of the required type was taken as the definition of the degree of asymptotic flatness required. I propose to give another definition for asymptotic flatness here, which is more evidently geometrical in character, but which is, in essentials, equivalent to the Bondi–Sachs definition. The idea involves the introduction of a *boundary* to space-time $\mathcal{M}$ whose points constitute future and past end-points to each null geodesic in $\mathcal{M}$. (We have just seen that the Bondi–Sachs mass is a 2-surface integral over "future end-points" to null geodesics, namely, the null geodesics which generate $\mathcal{N}_i$.) It turns out, somewhat miraculously, that Bondi–Sachs asymptotic flatness finds expression in the existence of a *smooth conformal structure* [74, 76] for the space-time with boundary. Having such a well-defined structure at infinity, we can then perform calculations by treating infinity as though it were a *local* structure and awkward asymptotic limits are thereby avoided.

Let us start by considering the nature of conformal infinity for Minkowski space-time $\mathscr{M}$. Choose null polar coordinates u, r, θ, φ related to the usual Minkowskian x^0, x^1, x^2, x^3 by

$$u = x^0 - r \qquad v = x^0 + r \tag{8.4}$$

with $r = [(x^1)^2 + (x^2)^2 + (x^3)^2]^{1/2}$ and

$$x^1 = r\cos\theta\cos\varphi \qquad x^2 = r\sin\theta\cos\varphi \qquad x^3 = r\sin\varphi. \tag{8.5}$$

Then

$$ds^2 = du\,dv - \tfrac{1}{4}(u - v)^2(d\theta^2 + \sin^2\theta\,d\varphi^2) \tag{8.6}$$

$(u \leqslant v)$. Let us choose new coordinates p, q with

$$v = \tan p \qquad u = \tan q \tag{8.7}$$

$(-\pi/2 \leqslant q \leqslant p \leqslant \pi/2$; see Fig. 15) so that points at infinity are assigned finite

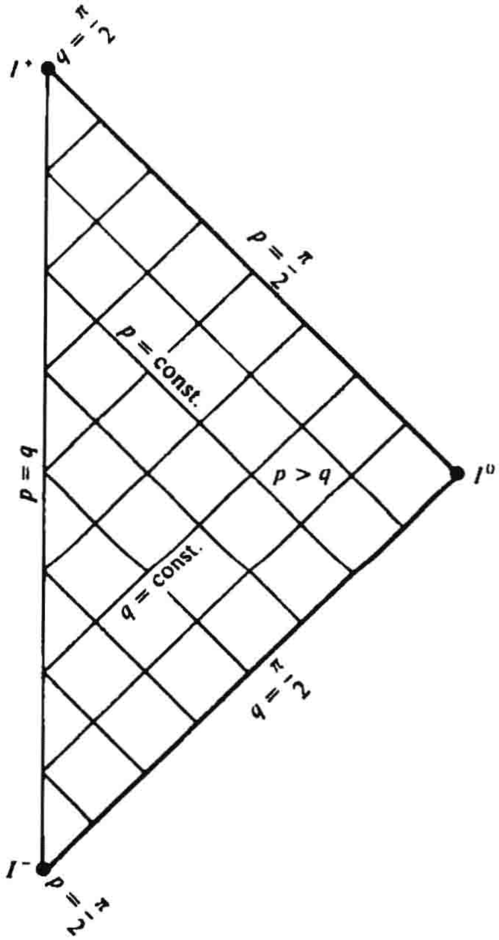


FIGURE 15. The range of p and q. To construct $\mathscr{M}$, rotate about $p = q$, so each point $p > q$ describes an S^2 (but keep I^0 fixed!).

p, q coordinate values. The metric (8.6) ceases to be meaningful at such values ($p = \pi/2$ or $q = -\pi/2$) but if we pass to a *conformally related* metric

$$d\hat{s} = \Omega\, ds \tag{8.8}$$

where we choose

$$\Omega = (1 + u^2)^{-1/2}(1 + v^2)^{-1/2} \tag{8.9}$$

then (8.6) becomes

$$d\hat{s}^2 = dp\, dq - \tfrac{1}{4} \sin^2(p - q)(d\theta^2 + \sin^2 \theta\, d\varphi^2) \tag{8.10}$$

The metric (8.10) is perfectly regular at $p = \pi/2$ and at $q = -\pi/2$ (except that at $p = q$ or $p - q = \pi$, we have removable coordinate singularities). Thus, we have a well-defined conformal structure on a manifold ($-\pi/2 \leqslant q \leqslant p \leqslant \pi/2$) with *boundary*[21] ($q = -\pi/2$ or $p = \pi/2$) whose interior ($-\pi/2 < q \leqslant p < \pi/2$) is identical in conformal structure with Minkowski space-time. Note that $q =$ const. and $p =$ const. are null hypersurfaces (cones) and are therefore generated by null geodesics. This then applies, in particular, to the boundary hypersurfaces.

The metric (8.10) is, in fact, that of *Einstein's static universe* $\mathscr{E}$. If we allow $0 \leqslant p - q \leqslant \pi$, with $p + q$ unrestricted (and with the usual range for the spherical polar coordinates θ, φ), then (8.10) refers to a "cylinder" $\mathscr{E} \cong S^3 \times E^1$, it being a simple matter to rewrite (8.10) in the form $d\hat{s}^2 = dT^2 - d\Sigma^2$, with $d\Sigma^2$ standing for the usual metric on a 3-sphere. (The regions of coordinate singularity in (8.10) can now be covered by applying a rotation in S^3 to the coordinates p, q, θ, φ.) The portion of $\mathscr{E}$ which corresponds to $\mathscr{M}$ can be described geometrically as follows (see Fig. 16). Choose a point I^- on $\mathscr{E}$ (this will be $p = q = -\pi/2$). The null geodesics on $\mathscr{E}$ which have past end-point at I^- will generate the future null cone of I^-. These null geodesics will reach a first focal point which is diametrically opposite (with respect to S^3) to I^-. Call this point I^0 (this is to be $p = -q = \pi/2$).

The (open) null geodesic segments from I^- to I^0 sweep out a hypersurface region $\mathscr{I}^-$ (to be $-\pi/2 < p < \pi/2$, $q = -\pi/2$) which is bounded at the past by I^- and at the future by I^0. ($\mathscr{I}^-$ has topology $S^2 \times E^1$.) The

---

[21] Strictly speaking, we do not quite have a manifold with boundary here, because of "corners" on the boundary. These will be removed for the strict definition to be given shortly.

There is another version of compactified conformal Minkowski space-time in which the future boundary hypersurface is identified with the past. Then the whole structure becomes a compact conformal manifold without boundary ($\cong S^1 \times S^3$) (see [51, 92]; also [76, 79]).

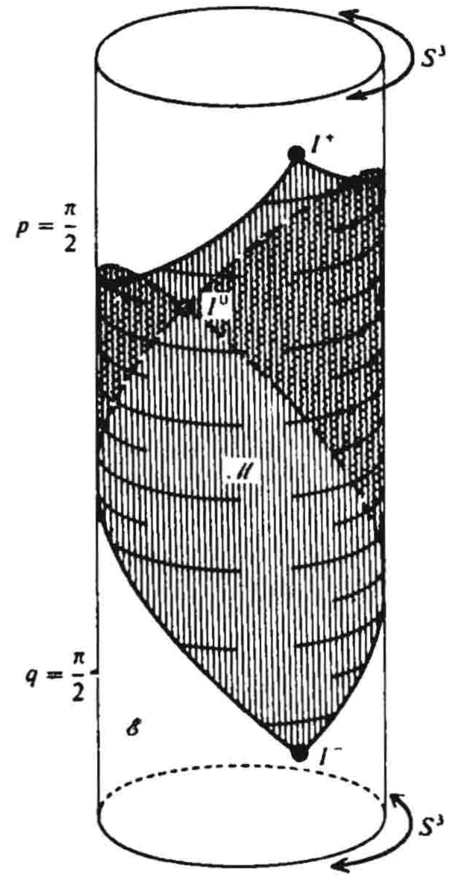


FIGURE 16. Conformal Minkowski space-time as a portion of the Einstein universe.

hypersurface $\mathscr{I}^-$ defines the past null cone at I^0. Continue these null geodesics into the future beyond I^0. They reach a second focal point I^+ (to be $p = q = \pi/2$), which is diametrically opposite to I^0 (with respect to S^3) and therefore lies directly "above" I^- in $\mathscr{E}$. The (open) null geodesic segments from I^0 to I^+ will sweep out a null hypersurface region $\mathscr{I}^+$ (to be $p = \pi/2$, $-\pi/2 < q < \pi/2$), again of topology $S^2 \times E^1$, and bounded by I^0 at the past and I^+ at the future. The part of $\mathscr{E}$ which may be identified with $\mathscr{M}$ (as regards conformal structure) is then just the set of points of $\mathscr{E}$ which lie "between" $\mathscr{I}^-$ and $\mathscr{I}^+$ (that is, the set of points which lie between I^- and I^+ on timelike curves from I^- to I^+). Thus $\mathscr{M}$ is an open subset of $\mathscr{E}$ and its boundary is $I^- \cup \mathscr{I}^- \cup I^0 \cup \mathscr{I}^+ \cup I^+$.

A useful way of picturing $\mathscr{M}$ is as the interior of two cones joined base to base (Fig. 17). Then the various parts of $\dot{\mathscr{M}}$ are represented as portions of these bounding cones. We must bear in mind that this picture is not *conformally* accurate, however. The inaccuracy is greatest near I^0, as emerges particularly since I^0 appears as an equatorial region whereas it

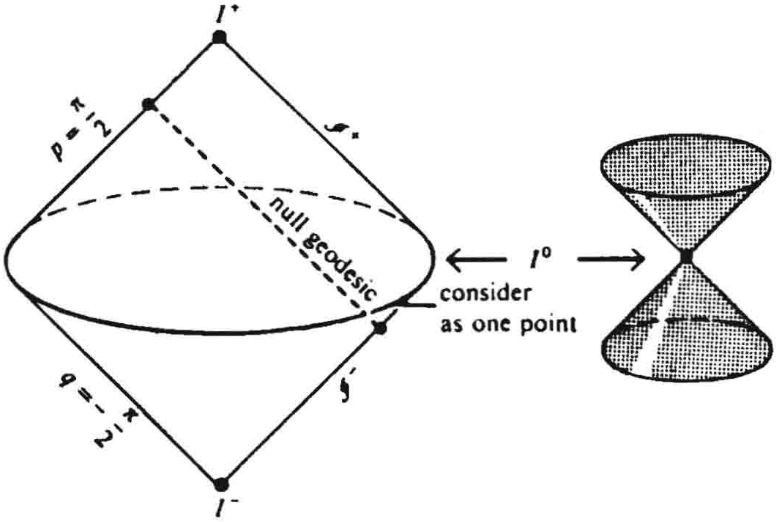


FIGURE 17. Any null geodesic in $\mathscr{M}$ attains two end-points in $\bar{\mathscr{M}}$, one on $\mathscr{I}^-$ and one on $\mathscr{I}^+$.

should be a point. (Of course, to be accurate, we should also have to think of the picture as four-dimensional.)

The interpretation of the points I^-, I^0, I^+ is as "past timelike infinity," "spacelike infinity," and "future timelike infinity." The hypersurfaces $\mathscr{I}^-$ and $\mathscr{I}^+$ are "past null infinity" and "future null infinity," respectively. The reason for labeling them as such becomes evident if we examine straight lines in $\mathscr{M}$ (straight according to the Minkowski metric ds). A timelike straight line becomes a curve from I^- to I^+; a spacelike straight line becomes a closed curve through I^0; a null straight line is a null geodesic originating at a point of $\mathscr{I}^-$ and terminating at a point of $\mathscr{I}^+$. In each case, the curve becomes "compactified" by the addition of the appropriate points at infinity.

The extension of $\mathscr{M}$ as a conformal manifold does not, of course, uniquely lead to the *entire* conformal Einstein universe. For, beyond the boundary $\dot{\mathscr{M}}$, we could bend the extended space as we wished. However, the closure $\bar{\mathscr{M}}$ of $\mathscr{M}$ in $\mathscr{E}$ *is* uniquely defined as a conformal manifold with boundary. For this to be quite precise, it is more logical to *delete* the points I^-, I^0, I^+ from the definition of $\bar{\mathscr{M}}$ and also from $\dot{\mathscr{M}} = \mathscr{I}$, since otherwise $\bar{\mathscr{M}}$ would not be a manifold with boundary at these three points. This will also fit in more accurately with the discussion of asymptotically flat space-times. (However, I shall not always be completely consistent in this respect, since it is useful to be able to talk about I^-, I^0, I^+.) There are various ways to see that this $\bar{\mathscr{M}}$ is conformally unique. Perhaps the most direct is to construct the points of $\mathscr{I}$ in terms of equivalence classes of null geodesics in $\mathscr{M}$ [an alternative way of defining the points of $\bar{\mathscr{M}}$ is in terms of event horizons and particle horizons (*cf.* Section 9). The nontrivial horizons can be used to define the points of $\dot{\mathscr{M}}$]. The concept of a ***null geodesic*** is a ***conformally invariant*** one, so such a construction is appropriate here. We have to decide when two null geodesics in $\mathscr{M}$ are to be deemed as intersecting on $\dot{\mathscr{M}}$. A method given by Geroch [34, 35] (see Dr. Geroch's lecture, Chapter VIII) can be used here to treat quite general space-times. In the present case, where $\mathscr{M}$ is Minkowskian, two null geodesics meet at the same point of $\mathscr{I}^+$ if and only if they belong (according to the ds metric) to the same ***null hyperplane*** in $\mathscr{M}$. If they are merely parallel in $\mathscr{M}$, then they just meet the same generator of I^+. For this particular space-time $\mathscr{M}$, two null geodesics meeting on $\mathscr{I}^+$ must also meet on $\mathscr{I}^-$ and *vice versa.* However, this is a very special property of Minkowski space-time and will not hold for general $\mathscr{M}$.

Let us pass to a slightly more general space-time $\mathscr{M}$, namely, the Schwarzschild solution for the metric of the exterior field of a spherically symmetric massive body. I shall use the Eddington–Finkelstein [27, 32] form of the metric with one retarded null coordinate u:

$$ds^2 = du^2[1-(2m/r)] + 2dr\,du - r^2(d\theta^2 - \sin^2\theta\,d\varphi^2) \tag{8.11}$$

($r > 0$). (This metric is discussed more completely in Section 10.) If we choose $d\hat{s} = \Omega\,ds$, with

$$\Omega = r^{-1} = l, \tag{8.12}$$

where l is to be a new coordinate which is finite for $r = \infty$, then we get

$$d\hat{s}^2 = du^2(l^2 - 2ml^3) - 2dl\,du - d\theta^2 - \sin^2\theta\,d\varphi^2 \tag{8.13}$$

($l \geqslant 0$). This is regular at $\mathscr{I}^+$, defined by $l = 0$, u finite. Here $\mathscr{I}^+$ is $S^2 \times E^1$ as before. To find $\mathscr{I}^-$, we re-express (8.11) in terms of an advanced null coordinate $v = u + 2r + 4m \log(r - 2m)$:

$$ds^2 = dv^2\left(1 - \frac{2m}{r}\right) - 2dr\,dv - r^2(d\theta^2 - \sin^2\theta\,d\varphi^2) \tag{8.14}$$

($r > 0$) and again use (8.12). This gives,

$$d\hat{s}^2 = dv^2(l^2 - 2ml^3) + 2dl\,dv - d\theta^2 - \sin^2\theta\,d\varphi^2 \tag{8.14a}$$

($l \geqslant 0$), which is regular at $\mathscr{I}^- \cong S^2 \times E^1$, defined by $l = 0$. The only real difference between this case and the previous Minkowskian one, as regards the structure of infinity, lies in the fact that we do *not* get *regular* points I^-, I^0, I^+ here. It is not surprising that I^- and I^+ turn out to be singular, since the source becomes concentrated at these points, at the two ends of its history. But I^0 is also a singular point, having infinite conformal curvature (although a C^0 conformal metric can be assigned at I^0). Thus we omit I^0 and $I^\pm$ from the definition of $\bar{\mathscr{M}}$ (and from $\mathscr{I} = \mathscr{I}^- \cup \mathscr{I}^+$).

The form of the Schwarzschild metric given in (8.11), (8.14) is a special case of

$$ds^2 = r^{-2} A dr^2 - 2B_i \, dx^i dr + r^2 C_{ij} \, dx^i \, dx^j \tag{8.15}$$

(i, j = 1, 2, 3), where A, B_i, C_{ij} are functions of x^1, x^2, x^3, and r, such that if we put $x^0 = r^{-1}$, then these functions are to be sufficiently differentiable (say C^3) as functions of x^a ($a = 0, 1, 2, 3$) at and in the neighborhood of the hypersurface $\mathscr{I}$ defined by $x^0 = 0$. If we assume that the relevant determinant in A, B_i, C_{ij} does not vanish, then clearly the metric $d\hat{s} = \Omega \, ds$, with $\Omega = r^{-1}$ is regular (C^3) on $\mathscr{I}$. The metric (8.15) includes all metrics of the Bondi–Sachs type, so a regularity assumption for $\mathscr{I}$ seems a not unreasonable one to impose if we wish to study asymptotically flat space-times and allow the possibility of gravitational radiation. In fact, with no field equations imposed to restrict R_{ab}, (8.15) is more general than the Bondi–Sachs metrics. In particular, (8.15) includes the de Sitter space-time and also asymptotically de Sitter space-times. These are cases which can arise if we have a cosmological term present in Einstein's vacuum equations.

The choice of Ω for (8.15) which is required to make the metric ds regular at $\mathscr{I}$, has the important property that its gradient at $\mathscr{I}$ does not vanish ($\partial\Omega/\partial x^a = (1, 0, 0, 0)$ at $\mathscr{I}$) and thus defines the normal direction to $\mathscr{I}$ there. It is convenient to supplement an assumption of regularity at $\mathscr{I}$ with the condition that Ω has nonvanishing gradient at $\mathscr{I}$, but under certain circumstances this condition can be deduced from the regularity assumption.

Let us be more precise as to the conditions to be satisfied. We require that our space-time $\mathscr{M}$, with metric ds^2, be extendible to a conformal manifold with boundary $\bar{\mathscr{M}} \supset \mathscr{M}$ (int $\bar{\mathscr{M}} = \mathscr{M}$; $\dot{\mathscr{M}} = \bar{\mathscr{M}} - \mathscr{M}$) such that

> THERE EXISTS A SMOOTH (SAY C^3, AT LEAST) REAL-VALUED FUNCTION $\Omega(\geqslant 0)$ ON $\bar{\mathscr{M}}$ AND A SMOOTH PSEUDO-RIEMANNIAN METRIC $d\hat{s}^2$ ON $\bar{\mathscr{M}}$ (CONSISTENT WITH ITS CONFORMAL STRUCTURE) SUCH THAT $ds^2 = \Omega^2 \, d\hat{s}^2$ ON $\mathscr{M}$ (8.16)
>
> ON $\dot{\mathscr{M}}$, WE HAVE $\Omega = 0$, $\hat{\nabla}_a \Omega \neq 0$ (8.17)
>
> EVERY NULL GEODESIC IN $\mathscr{M}$ HAS TWO END-POINTS ON $\dot{\mathscr{M}}$ (8.18)

If such an $\bar{\mathscr{M}}$ exists, we call $\mathscr{M}$ *asymptotically simple*. Then $\bar{\mathscr{M}}$ is unique (as follows, for example, by Geroch's construction of $\dot{\mathscr{M}}$ in terms of null geodesics [34]). We write $\dot{\mathscr{M}} = \mathscr{I}$, for the points at infinity for $\mathscr{M}$.

The condition (8.18) is to ensure that the whole of infinity for $\mathscr{M}$ has been included. It is generally difficult to verify in practice, however, and is not even satisfied by some space-times that one might like to think of as being asymptotically *flat* but which contain bound null orbits (that is, null geodesics which do not escape to infinity such as are found at $r = 3m$ in the Schwarzschild solution). In order to cover this type of possibility, let us call a space-time $\mathscr{M}$ *weakly asymptotically simple* if an asymptotically simple $\mathscr{M}_0$ exists, such that for some open subset $\mathscr{K}$ of $\bar{\mathscr{M}}_0$, with $\dot{\mathscr{M}}_0 \subset \mathscr{K}$, the region $\mathscr{M}_0 \cap \mathscr{K}$ (with metric induced from $\mathscr{M}_0$) is isometric with a subset of $\mathscr{M}$. That is to say, a weakly asymptotically simple space-time possesses the conformal infinity $\mathscr{I} = \dot{\mathscr{M}}_0$ of an asymptotically simple space-time but it may possess other "infinities" as well.

To obtain the connection with Bondi–Sachs asymptotic flatness, let us suppose that $\mathscr{M}$ is weakly asymptotically simple with conformal infinity $\mathscr{I}$. Assuming Einstein's equations (7.1), it can be shown that, provided T_{ab} does not approach some nonzero multiple of g_{ab} near $\mathscr{I}$, then

$$\mathscr{I} \text{ IS SPACELIKE, TIMELIKE, OR NULL, ACCORDING AS THE COSMOLOGICAL CONSTANT } \lambda \text{ IS POSITIVE, NEGATIVE, OR ZERO.} \tag{8.19}$$

[Also, if $\lambda \neq 0$, then we do not need to assume the condition $\hat{\nabla}_a \Omega \neq 0$ of (8.17) but we can deduce it.] If $\mathscr{I}$ is spacelike or null, then $\mathscr{I}$ is the disjoint union of two null hypersurfaces $\mathscr{I}^-$, $\mathscr{I}^+$. The points of $\mathscr{I}^-$ are distinguished from the points of $\mathscr{I}^+$ by the fact that their future null cones, rather than their past null cones, lie in $\bar{\mathscr{M}}$. We are interested here in asymptotic flatness, rather than cosmology, so let us assume that $\mathscr{I}$ is *null*. Then it follows[22] (provided $\mathscr{M}_0$, in addition to being time- and space-orientable, is free of closed timelike curves—possibly an unnecessary assumption) that

$$\mathscr{I}^- \cong \mathscr{I}^+ \cong S^2 \times E^1 \tag{8.20}$$

the E^1's being the null generators of $\mathscr{I}^\pm$. The result (8.20) shows that when $\mathscr{I}$ is null, it fairly closely resembles the infinity of Minkowski space-time.

To proceed further, we need stronger conditions on the way that T_{ab} behaves near $\mathscr{I}$. For simplicity of the discussion, let us assume, for the moment, that $T_{ab} = 0$ (and $\lambda = 0$) in the neighborhood of $\mathscr{I}$. Then more can be said. In fact,

$$\hat{\nabla}_{A(A'} \hat{\nabla}_{B')}{}^{B} \Omega = 0 \text{ AT } \mathscr{I} \tag{8.21}$$

and as a consequence of (8.21) and (8.20), cf. [76],

$$\hat{\Psi}_{ABCD} = 0 \text{ AT } \mathscr{I} \tag{8.22}$$

[We can also obtain $\hat{\nabla}\Omega \neq 0$ at $\mathscr{I}$, rather than assume it in (8.17).] On the basis of (8.22) it is possible to deduce the *peeling-off property* for principal null directions that was mentioned in Section 7. The arguments involved in this will be indicated next.

In spinor form, the Bianchi identities (8.2) become

$$\nabla^{A}_{P'} \Psi_{ABCD} = \nabla^{Q'}_{B} \Phi_{CDQ'P'} - 2\varepsilon_{B(C} \nabla_{D)P'} \Lambda \tag{8.23}$$

Thus, with $T_{ab} = 0$, we get

$$\nabla^{AP'} \Psi_{ABCD} = 0 \tag{8.24}$$

---

[22] An outline of an argument for obtaining this result is given in an appendix in [76]. It would seem that there must be a more satisfactory way of obtaining the result, however.

Equation (8.24) is of interest, because it is the curved-space version of the zero rest-mass free-field equation in the special case of spin $s = 2$. For general nonzero spins $s = \frac{1}{2}, 1, 1\frac{1}{2}, 2, \ldots$ we would have

$$\nabla^{AP'}\varphi_{AB\ldots L} = 0 \tag{8.25}$$

where

$$\varphi_{AB\ldots L} = \varphi_{(AB\ldots L)} \tag{8.26}$$

is a spinor with $2s$ indices. For $s = \frac{1}{2}$, Eq. (8.25) is the Weyl neutrino equation (zero-mass Dirac equation). For $s = 1$, (8.25) is the spinor version of Maxwell's free-space equations

$$\nabla_{[a}F_{bc]} = 0 \quad \nabla^{a}F_{ab} = 0 \tag{8.27}$$

with the Maxwell field tensor

$$F_{ab} = -F_{ba} = \varphi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\bar{\varphi}_{A'B'} \tag{8.28}$$

In flat space-time, (8.25) is a very satisfactory equation, the solutions having the same number of "degrees of freedom" (namely two) for *any* value of s as they do in the two cases $s = \frac{1}{2}, 1$. However, in curved space-time we have the consistency condition [16, 84],

$$\begin{aligned}(2-2s)\Psi^{ABM}{}_{(C}\varphi_{D\ldots L)ABM} &= \nabla^{(B}_{P'}\nabla^{A)P'}\varphi_{ABCD\ldots L} \\ &= \nabla^{B}_{P'}\nabla^{AP'}\varphi_{ABCD\ldots L} = 0\end{aligned} \tag{8.29}$$

so that if $s \geqslant \frac{3}{2}$, (8.25) implies an interconnection between $\varphi_{AB\ldots L}$ and the conformal curvature spinor Ψ_{ABCD}. In the case $\varphi_{ABCD} = \Psi_{ABCD}$, (8.29) is, however, automatically satisfied because of symmetry considerations. We leave aside the general question of higher-spin field equations here and interpret (8.25) only in the cases $s = \frac{1}{2}$ (neutrinos) $s = 1$ (electromagnetism) and $s = 2$ (gravitation).

An important feature of (8.25) is that it is *conformally invariant* with

$$\hat{\varphi}_{AB\ldots L} = \Omega^{-1}\varphi_{AB\ldots L} \tag{8.30}$$

We require various conformal transformation formulas. From $d\hat{s} = \Omega\, ds$ we get

$$\hat{g}_{ab} = \Omega^{2}g_{ab} \qquad \hat{g}^{ab} = \Omega^{-2}g^{ab} \tag{8.31}$$

so we can set[23]

---

[23] A somewhat different convention has been used here for treating spinors under conformal transformation than that used in [76], resulting in different transformation formulas. This difference is forced upon us by the attitude adopted in Section 3. In fact, the formulas obtained here are a little simpler than those of [76] and fractional powers of Ω are avoided.

$$\hat{\varepsilon}_{AB} = \Omega \varepsilon_{AB} \qquad \hat{\varepsilon}_{A'B'} = \Omega \varepsilon_{A'B'} \qquad \hat{\varepsilon}^{AB} = \Omega^{-1} \varepsilon^{AB} \qquad \hat{\varepsilon}^{A'B'} = \Omega^{-1} \varepsilon^{A'B'} \tag{8.32}$$

For a scalar χ, we have

$$\hat{\nabla}_a \chi = \nabla_a \chi \tag{8.33}$$

For indexed quantities we have, for example,

$$\begin{aligned} &\hat{\nabla}_{AA'} \xi_B = \nabla_{AA'} \xi_B - \Upsilon_{BA'} \xi_A \qquad \hat{\nabla}_{AA'} \eta_{B'} = \nabla_{AA'} \eta_{B'} - \Upsilon_{AB'} \eta_{A'} \\ &\hat{\nabla}_{AA'} \xi^B = \nabla_{AA'} \xi^B + \varepsilon_A{}^B \Upsilon_{CA'} \xi^C \qquad \hat{\nabla}_{AA'} \eta^{B'} = \nabla_{AA'} \eta^{B'} + \varepsilon_{A'}{}^{B'} \Upsilon_{AC'} \eta^{C'} \end{aligned} \tag{8.34}$$

where

$$\Upsilon_{AA'} = \Omega^{-1} \nabla_{AA'} \Omega \tag{8.35}$$

To treat a quantity with more than one index we simply treat it as though it were a product, according to the Leibnitz rule. That is to say, we get one term in $\Upsilon_{XY'}$ for each index. To verify the correctness of (8.34) we need only show that the axioms (4.50)–(4.58) are satisfied.

From (8.34), we get that (8.30) substituted in (8.25) yields

$$\hat{\nabla}^{AP'} \hat{\phi}_{AB \ldots L} = 0 \tag{8.36}$$

which is the required conformal invariance. However, when applying this to the gravitational field, we have to be careful. The Weyl tensor has the conformal invariance property $\hat{C}^a{}_{bcd} = C^a{}_{bcd}$, that is, $\hat{C}_{abcd} = \Omega^2 C_{abcd}$. Hence

$$\hat{\Psi}_{ABCD} = \Psi_{ABCD} \tag{8.37}$$

Thus, if according to the physical metric *ds* we set

$$\varphi_{ABCD} = \Psi_{ABCD} \tag{8.38}$$

then we have [by (8.30) and (8.37)]

$$\hat{\phi}_{ABCD} = \Omega^{-1} \hat{\Psi}_{ABCD} \tag{8.39}$$

It is the field $\hat{\phi}_{ABCD}$ which we identify with the gravitational field when we examine $\hat{\mathscr{M}}$ according to its metric $d\hat{s}$.

The significance of this lies in the behavior of the field on $\mathscr{I}$ and the relation of this to the peeling-off property. Recall, from (8.22), that Ψ_{ABCD} *vanishes* on $\mathscr{I}$. From (8.39) and smoothness we obtain:

$$\hat{\phi}_{ABCD} \text{ IS CONTINUOUS AT } \mathscr{I} \tag{8.40}$$

Since it costs us no more at this point, let us derive the peeling-off property for *general* spin s on the basis of the assumption that $\hat{\phi}_{AB \ldots L}$ is continuous at $\mathscr{I}$.

We require a spin frame $o^A = \varepsilon_0{}^A$, $\iota^A = \varepsilon_1{}^A$. For the $d\hat{s}$ metric set

$$\hat{o}_A = o_A \qquad \iota_A = \Omega \iota_A \qquad \hat{o}^A = \Omega^{-1} o^A \qquad \hat{\iota}^A = \iota^A \tag{8.41}$$

This is consistent with (8.32) and (4.27). The reason for the asymmetrical choice (8.41) is that we wish to arrange o_A to have its direction pointing along a null geodesic η, with o_A, ι_A parallelly propagated (according to ds) along η:

$$\nabla_{00'} o^A = 0 \qquad \nabla_{00'} \iota^A = 0 \tag{8.42}$$

Then (8.34) and (8.41) yield

$$\hat{\nabla}_{00'} \hat{o}^A = 0 \qquad \hat{\nabla}_{00'} \hat{\imath}^A = \Omega^{-1} \hat{o}^A \hat{\nabla}_{10'} \Omega \tag{8.43}$$

Thus o^A is parallelly propagated according to $d\hat{s}$ along η and attains a *well-defined value* at $Q^{\pm} = \eta \cap \mathscr{I}^{\pm}$. In fact $\hat{\imath}^A$ also has a well-defined limiting value at $Q^{\pm}$ and points along $\mathscr{I}^{\pm}$ at $Q^{\pm}$ (so $\hat{\nabla}_{10'} \Omega = 0$ at $\mathscr{I}$; cf. [76]). The affine parameter r on η (affine according to ds) defined by

$$\nabla_{00'} r = 1 \tag{8.44}$$

has the asymptotic behavior

$$r \sim \Omega^{-1} \tag{8.45}$$

Now, by hypothesis, the components $\hat{\phi}_{\mathfrak{AB}\dots\mathfrak{L}}$, according to the $\hat{o}^A$, $\hat{\imath}^A$ spin frame, are all continuous at $Q^{\pm}$. Substituting (8.30), (8.41), and (8.45) we interpret this in terms of the physical quantities as

$$\lim_{r \to \pm\infty} \{ r^{2s+1-i} \varphi_{(i)} \} \quad \text{EXISTS} \quad (i = 0, 1, \dots, 2s) \tag{8.46}$$

where $\varphi_{(i)}$ stands for $\varphi_{00\dots01\dots1}$ with i "ones" and $2s - i$ "zeros." Equation (8.46) tells us that the r^{-k} part of $\varphi_{AB\dots L}$ along η satisfies

$$\varphi_{AB\dots DE\dots L} o^A o^B \dots o^D = 0 \tag{8.47}$$

($k = 1, 2, \dots, 2s$) where k spinors o^A appear in (8.47). Condition (8.47) is just the condition for at least $2s - k + 1$ principal null direction of $\varphi_{A\dots L}$ to coincide in the direction of η. This, then, is the peeling-off property.

For the case of gravitation, this peeling-off property is a characteristic feature of Bondi–Sachs asymptotic flatness. It is possible also, to relate more directly the precise type of coordinates used in the Bondi–Sachs approach to the conformal structure that arises in the present formalism. This has been done explicitly by Tamburino and Winicour [104]. In the present approach we may thus *define asymptotic flatness* for $\mathscr{M}$ just to be *weak asymptotic simplicity*, in the case when $\lambda = 0$ and $T_{ab} = 0$ near $\mathscr{I}$. If T_{ab} only approaches zero in some way toward infinity, we may need further conditions, such as (8.21), (8.22) in addition to weak asymptotic simplicity. In the case of the Einstein–Maxwell equations holding near $\mathscr{I}$, then (except for a possibility of some pathological exceptional cases) such extra conditions are not needed, and peeling-off can be deduced for both gravitation and electromagnetism.

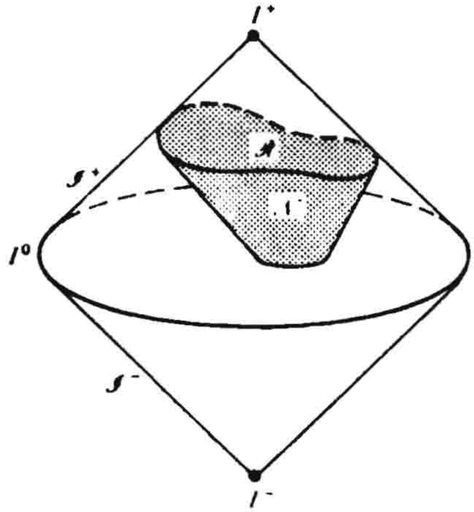


FIGURE 18. The Bondi–Sachs energy-momentum is an integral over $\mathscr{R} = \mathscr{N} \cap \mathscr{I}^+$.

Let us return, briefly, to the question of energy-momentum with which we started. If the outgoing null hypersurface $\mathscr{N}$ meets $\mathscr{I}^+$ in the 2-surface $\mathscr{R}$ (Fig. 18), then we can perform an integral over $\mathscr{R}$ to obtain the total energy-momentum intercepted by any hypersurface spanning $\mathscr{R}$. If we choose Ω so that the metric of $\mathscr{R}$ is that of a unit sphere, then we integrate

$$\hat{\sigma} N - \Psi_2 \tag{8.48}$$

over $\mathscr{R}$ to get the energy component (associated with this choice of Ω) and we integrate (8.48) with a suitable weighting factor to get the momentum components. Here

$$\Psi_2 = \phi_{0011} \tag{8.49}$$

on $\mathscr{R}$, so Ψ_2 comes from the r^{-3} part of the Weyl curvature in the peeling-off. The quantity $\hat{\sigma}$ is the shear $\hat{\Gamma}_{0001}$, of $\mathscr{N}$ at $\mathscr{R}$ (with $\hat{o}^A$ and $\hat{\imath}^A$ describing the null direction in $\mathscr{N}$ and $\mathscr{I}$, respectively). The quantity $\mathscr{N}$ is the Bondi–Sachs "news function" given here by

$$N = \hat{\Phi}_{110'0'} \tag{8.50}$$

Although this is a Ricci tensor component, it retains a certain conformal covariance by virtue of the above restriction on Ω. The whole energy-momentum turns out to have the correct transformation behavior for an "asymptotic 4-vector."

The Bondi–Sachs mass-loss formula finds expression in the fact that if we repeat the mass integral with a new null hypersurface lying to the future of the original one, the mass value so obtained is always less than or equal to the earlier mass value. The mass carried by the gravitational waves turns out to be the integral of $N\bar{N}$ over the part of $\mathscr{I}^+$ lying between the two null hypersurfaces. In fact $\bar{N}$ may be described as an integral, along the generators of $\mathscr{I}^+$ of $\Psi_4 = \hat{\phi}_{1111}$. [This follows from (8.50), in fact.] In the peeling-off, Ψ_4 is the r^{-1} part of the gravitational field and so may be thought of as the outgoing *radiation field.* "News function" is therefore essentially just the "time-integral of the outgoing gravitational radiation field." It turns out that the Bondi–Sachs mass loss formula is closely related to the positive focusing property (7.39) on a null line γ which lies just inside $\mathscr{I}^+$. Here N takes the place of σ in (7.39) and Ψ_4 takes the place of Ψ. (See [78].)

It may be remarked that the mass loss occurs *irrespective* of the presence of *incoming* radiation (that is, analog of Ψ_4 on $\mathscr{I}^-$). Such asymptotic flatness assumptions at $\mathscr{I}^+$ do *not* exclude incoming waves (as seems to have been believed at one time), although they forbid a too large concentration of incoming waves just inside $\mathscr{I}^+$. To see the mass *gain* due to incoming radiation, we should have to turn the above construction upside-down and perform our calculations on $\mathscr{I}^-$ instead.

As a final remark concerning energy in general relativity, we see that while the (future-measured) mass always decreases, there is no guarantee that the mass measurement never becomes negative! In the cases in which all sources finally are radiated away to $\mathscr{I}^+$ as zero-rest-mass fields, we can say from the Bondi–Sachs formula that the total mass *was* always positive.[24] But to my knowledge, it is an unsolved problem whether the total mass is positive in *general* situations where we just assume a suitable physically reasonable inequality on T_{ab}.

9 HORIZONS

When examining specific cosmological models, it is sometimes useful to represent infinity conformally, even though the space-time may not be (weakly) asymptotically simple. For example, in the usual Robertson–Walker "big bang" models, we can represent the initial infinite curvature singularity as a nonsingular hypersurface boundary in the past, but only if we allow $\Omega = \infty$ at this region. The nature of such boundary hypersurfaces is closely related to the question of four types of *horizon* which may occur in cosmology: the *event* horizon or *particle* horizon [88] and future or past *Cauchy* horizons [40, 42]. These will be defined precisely in a moment.

I shall first require some definitions. By a *curve* I mean *the image* of a (noncritical) C^1-map f of a real interval of nonzero length, where I shall demand that every point of $\mathscr{M}$ has a neighborhood $\mathscr{Q}$ for which $f^{-1}\,\bar{\mathscr{Q}}$ has no noncompact connected component. Thus *end-points* are required (at which the curve would have to be smooth, and therefore extendible) except where the curve continues indefinitely in one direction or the other. (Otherwise we could refer to an open-ended curve.) If the map f itself is required, this is a *parametrized* curve. A curve is timelike [resp. nonspacelike] if the tangent vectors of the map f are all timelike [resp. timelike or null]. A nonspacelike curve has a natural orientation, called a *future orientation*, which is induced by the time-orientation of $\mathscr{M}$.

---

[24] Brill [14] obtained some partial results concerning positiveness of total mass of a system as measured on a spacelike hypersurface. A promising new approach to this problem has recently been given by Brill and Deser [14a].

If P and Q are two points of $\mathscr{M}$, then (see [50]) the notation $P \ll Q$ is used if there is a *timelike curve in* $\mathscr{M}$ *with past end-point* P *and future end-point* Q. The notation $P \prec Q$ is used if either $P = Q$ or there is a nonspacelike curve from P to Q. If $P \prec Q$ and $Q \ll R$ or if $P \ll Q$ and $Q \prec R$, then $P \ll R$. If $P \prec Q$ (with $P \neq Q$), then either $P \ll Q$ or there is a null geodesic from P to Q (or both). The chronological *future* and *past* of P, respectively, are

$$I_+(P) = \{X : P \ll X\} \qquad I_-(P) = \{X : X \ll P\} \tag{9.1}$$

These are *open* sets, as is easily seen. On the other hand, the sets

$$J_+(P) = \{X : P \prec X\} \qquad J_-(P) = \{X : X \prec P\} \tag{9.2}$$

are not necessarily closed. For example, let $\mathscr{M}$ be Minkowski space with a point removed and let P be on the null cone of the removed point. For another example, consider a plane wave, *cf.* (9.23)–(9.31). If $\mathscr{X}$ is any subset of $\mathscr{M}$, the chronological future and past of $\mathscr{X}$ are denoted, respectively, by

$$I_+[\mathscr{X}] = \bigcup_{P \in \mathscr{X}} I_+(P) \qquad I_-[\mathscr{X}] = \bigcup_{P \in \mathscr{X}} I_-(P) \tag{9.3}$$

Again these are open sets. The *closures* of these sets are

$$\bar{I}_+[\mathscr{X}] = \{X : I_+(X) \subset I_+[\mathscr{X}]\} \qquad \bar{I}_-[\mathscr{X}] = \{X : I_-(X) \subset I_+[\mathscr{X}]\} \tag{9.4}$$

The proof of (9.4) is easy and will be omitted here. From (9.4), we obtain[25] the *boundaries* of $I_+[\mathscr{X}]$:

$$\begin{aligned} \dot{I}_+[\mathscr{X}] &= \{X : I_+(X) \subset I_+[\mathscr{X}],\ X \notin I_+[\mathscr{X}]\} \\ \dot{I}_-[\mathscr{X}] &= \{X : I_-(X) \subset I_-[\mathscr{X}],\ X \notin I_-[\mathscr{X}]\} \end{aligned} \tag{9.5}$$

A set of the form $\dot{I}_+[\mathscr{X}]$ will be called a *semispacelike boundary* or SSB. In fact, this is *time-symmetrical* since

$$\dot{I}_+[\mathscr{X}] = \dot{I}_-[\mathscr{L}] \tag{9.6}$$

---

[25] The notation $\mathscr{A} \subset \mathscr{B}$ means just that $\mathscr{A}$ is a subset of $\mathscr{B}$, not necessarily a proper subset (so $\mathscr{A} \subset \mathscr{A}$).

where $\mathscr{L}$ is the complement in $\mathscr{M}$ of $\dot{I}_+[\mathscr{K}]$. Conversely, (9.6) still holds if it is $\mathscr{L}$ that is an arbitrary subset of $\mathscr{M}$ and $\mathscr{K}$ is, for example, the complement of $\dot{I}_-[\mathscr{L}]$. A *semispacelike set*, in general, is a subset of $\mathscr{M}$ no two points X, Y of which satisfy $X \ll Y$. Clearly an SSB is semispacelike. Although SSB's need not be smooth subsets of $\mathscr{M}$, they have certain nice properties. These will play an important part in the "singularity theorems" of Sections 10 and 11. For the moment, we simply note that:

> ANY SSB IS A C^0 THREE-DIMENSIONAL SUBMANIFOLD OF $\mathscr{M}$ (WITHOUT BOUNDARY; AND IS A CLOSED SET) (9.7)

To see that an SSB is locally homeomorphic to E^3, we examine a small normal coordinate neighborhood (x^a) of any point P of the SSB (with $\nabla_a x^0$ timelike and $\nabla_a x^i$ spacelike, $i \neq 0$) any curve $x^i = \text{const.}$ $(i \neq 0)$ which is near enough to P intersects both $I_+(P)$ and $I_-(P)$. It therefore intersects the SSB at a point which must be unique. This gives a local homeomorphism onto an open set of the $\mathbf{R}^3$ of (x^1, x^2, x^3) (Fig. 19).

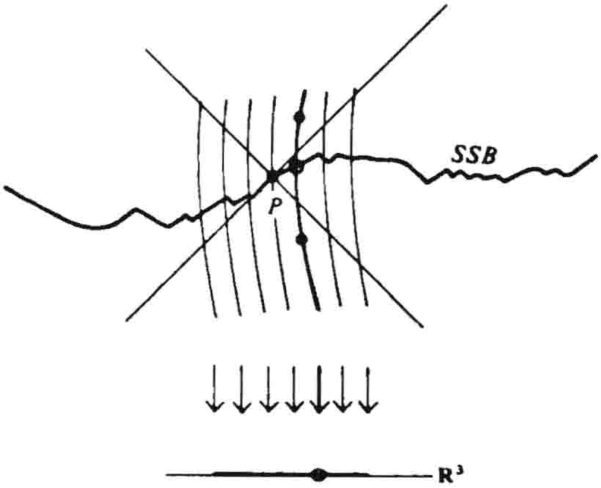


FIGURE 19. Any SSB is locally homeomorphic to E^3.

We are now in a position to define our four types of horizon. Let γ be any timelike curve in $\mathscr{M}$. Then (when nonvacuous)

$$\begin{aligned} &\text{EVENT HORIZON OF } \gamma = \dot{I}_-[\gamma] \\ &\text{PARTICLE HORIZON OF } \gamma = \dot{I}_+[\gamma] \end{aligned} \tag{9.7a}$$

If γ has a future end-point F, then its event horizon is simply $\dot{I}_-(F)$; if γ has a past end-point P, then its particle horizon is $\dot{I}_+(P)$. These cases will be called *trivial*. An example of a space-time with *no* nontrivial event or particle horizons is the Einstein static universe mentioned in Section 8. However, Minkowski space-time *does* possess nontrivial event and particle

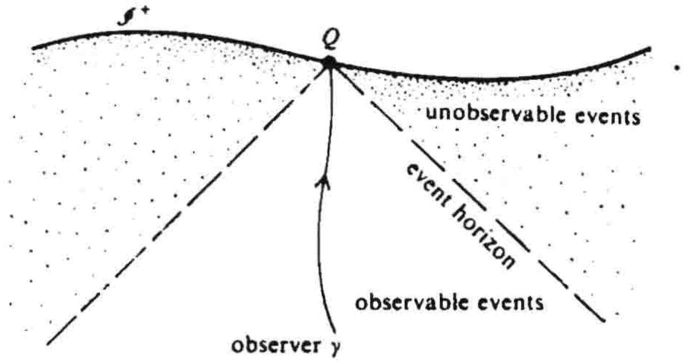


FIGURE 20. If $\mathscr{I}^+$ is spacelike, every timelike curve possesses an event horizon.

horizons, given, for example, when γ is a path of uniform acceleration. [See (2.8)–(2.10) and Fig. 6. If γ is $X = Y = 0$, $Z =$const. > 0, then $z = t$ is an event horizon of γ and $z = -t$ is a particle horizon of γ.] On the other hand, no timelike *geodesic* in Minkowski space-time has an event or particle horizon. Physically, the event horizon of γ separates those events which are observable by an observer whose world line is γ from those events unobservable by him. Similarly, the particle horizon of γ separates those events from which a particle with world line γ can be observed (in principle), from those events from which the particle cannot be so observed.

Examples of space-times with nontrivial event or particle horizons for *geodesic* curves γ, are obtained if $\mathscr{M}$ possesses a *spacelike* conformal infinity $\mathscr{I}^+$ or $\mathscr{I}^-$. (We allow $\Omega \neq 0$ or $\Omega = \infty$ at $\mathscr{I}$.) If $\mathscr{I}^+$ is spacelike, then any timelike geodesic γ (indeed, *any* inextendible timelike curve in $\mathscr{M}$) through a point of $\mathscr{M}$ near enough to $\mathscr{I}^+$, will hit $\mathscr{I}^+$ at some point Q in $\bar{\mathscr{M}}$. We can define $\dot{I}_-(Q)$ (from the conformal structure of $\bar{\mathscr{M}}$) and this will be the required event horizon $\dot{I}_-[\gamma]$ (Fig. 20). Exactly similarly, if $\mathscr{I}^-$ is spacelike, we can find timelike geodesics γ with nontrivial particle horizons (Fig. 21).

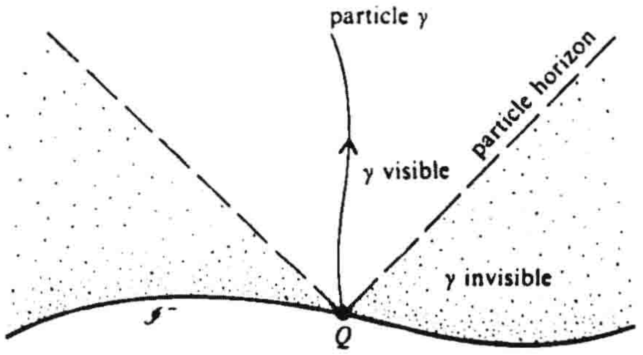


FIGURE 21. If $\mathscr{I}^-$ is spacelike, every timelike curve possesses a particle horizon.

To define the Cauchy horizons, let $\mathscr{S}$ be any closed semispacelike set. Define the *domains of dependence* $D_+(\mathscr{S})$, $D_-(\mathscr{S})$ as follows:

> $D_+(\mathscr{S})$ IS THE COMPLEMENT OF THE UNION OF ALL TIMELIKE CURVES WHICH ARE INEXTENDIBLE IN THE PAST AND WHICH DO NOT INTERSECT $\mathscr{S}$ (9.8)

with $D_-(\mathscr{S})$ as just the time-reverse of (9.8). Another way of putting this is

> $X \in D_+(\mathscr{S})$ IF AND ONLY IF EVERY TIMELIKE CURVE THROUGH X CAN BE EXTENDED INTO THE PAST UNTIL IT MEETS $\mathscr{S}$ (9.9)

and similarly for $D_-(\mathscr{S})$. [We have $\mathscr{S} \subset D_+(\mathscr{S})$.] It is readily seen that the past-inextendible curves not meeting $\mathscr{S}$ sweep out an open subset of $\mathscr{M}$. Thus, by (9.8)

> $D_+(\mathscr{S})$ AND $D_-(\mathscr{S})$ ARE CLOSED SETS (9.10)

The *Cauchy horizons* (Fig. 22) are the "future boundary" of $D_+(\mathscr{S})$ and the "past boundary" of $D_-(\mathscr{S})$:

$$\begin{aligned} H_+(\mathscr{S}) &= \{X\colon X \in D_+(\mathscr{S}),\ I_+(X) \cap D_+(\mathscr{S}) = \varnothing\} \\ H_-(\mathscr{S}) &= \{X\colon X \in D_-(\mathscr{S}),\ I_-(X) \cap D_-(\mathscr{S}) = \varnothing\} \end{aligned} \tag{9.11}$$

We have, easily,

> $H_+(\mathscr{S})$ AND $H_-(\mathscr{S})$ ARE SEMISPACELIKE CLOSED SETS (9.12)

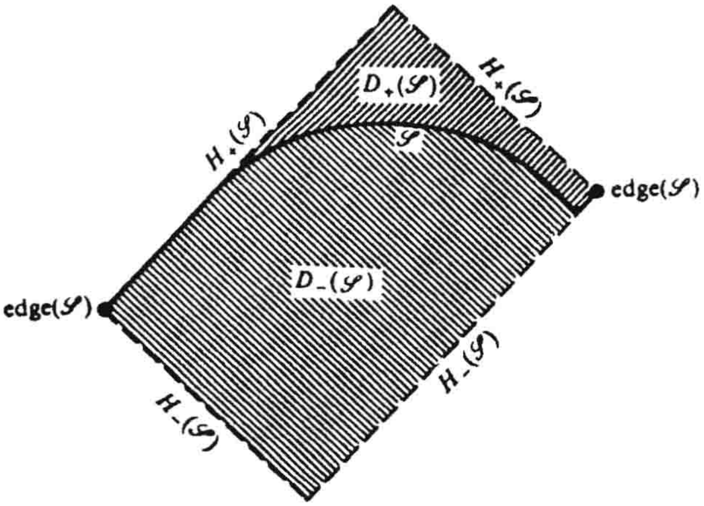


FIGURE 22. Domains of dependence and Cauchy horizons for a semispacelike set $\mathscr{S}$ (in a "trivial" case).

We can define the *edge* of a closed semispacelike set $\mathscr{S}$ as the set of points $P \in \mathscr{S}$ such that for any open set $\mathscr{Q}$ containing P, if $R \ll P \ll Q$ with $R, Q \in \mathscr{Q}$ then there is a timelike curve in $\mathscr{Q}$ from R to Q not meeting $\mathscr{S}$. We have

$$\text{edge}\,(\mathscr{S}) \subset H_+(\mathscr{S}) \cap H_-(\mathscr{S}) \tag{9.13}$$

The set edge ($\mathscr{S}$) is always closed. It consists of those points of $\mathscr{S}$ at which $\mathscr{S}$ is not locally homeomorphic to E^3. If edge $(\mathscr{S}) = \varnothing$, we can call $\mathscr{S}$ *edgeless*. If $\mathscr{S}$ is a smooth semispacelike set which is a *hypersurface* (without boundary) in $\mathscr{M}$ (so the tangent hyperplanes are everywhere spacelike or null), then $\overline{\mathscr{S}}$ is edgeless if and only if $\mathscr{S}$ is a *closed* subset of $\mathscr{M}$ (that is, $\mathscr{S}$ is properly embedded). As another example, any SSB is edgeless.

A Cauchy horizon (nonvacuous) will be called *nontrivial* if it is the $H_+(\mathscr{S})$ or $H_-(\mathscr{S})$ of some edgeless closed semispacelike set $\mathscr{S}$. Trivial Cauchy horizons are easily constructed (in space-times with no closed timelike curves) by just taking $\mathscr{S}$ to be a small closed portion of a spacelike hypersurface. The Einstein static universe contains no nontrivial Cauchy horizons. On the other hand, Minkowski space-time *does* contain nontrivial Cauchy horizons {for example, if $\mathscr{S}$ is $x^0 = x^1$, with usual Minkowskian coordinates, then $H_\pm(\mathscr{S}) = \mathscr{S}$; if $\mathscr{S}$ is $x^0 = -[1 + (x^1)^2 + (x^2)^2 + (x^3)^2]^{1/2}$, then $H_-(\mathscr{S}) = \varnothing$, but $H_+(\mathscr{S})$ is $x^0 = -[(x^1)^2 + (x^2)^2 + (x^3)^2]^{1/2}$}.

A *global Cauchy hypersurface* or GCH for $\mathscr{M}$ is a semispacelike set $\mathscr{S}$ (normally a smooth hypersurface) for which $H_+(\mathscr{S})$ and $H_-(\mathscr{S})$ are both empty. That is to say, $\mathscr{S}$ meets every inextendible timelike curve in $\mathscr{M}$. An equivalent condition on a semispacelike set $\mathscr{S}$ for it to be a GCH is that it should *intersect every inextendible null geodesic in* $\mathscr{M}$ *in a nonvacuous compact set*. This will follow from the later considerations concerning the structure of Cauchy horizons. It is not hard to see that the intersection of a null geodesic with an SSB must be *connected*. One property of a space-time $\mathscr{M}$, containing a GCH $\mathscr{S}$, which is easy to establish is that it is topologically $\mathscr{S} \times E^1$, where the E^1's can be represented as timelike curves. It is less easy to establish, but still true,[26] that all the $\mathscr{S}$'s of the product can also be made to be GCH's for $\mathscr{M}$. It also turns out that the existence of a GCH is equivalent to Leray's "global hyperbolicity" condition.[27]

It is clear that the usual "constant time" sections of the Einstein static universe and of Minkowski space-time are GCH's. Among the examples of space-times which do *not* contain any GCH are all those which possess a *timelike* conformal infinity $\mathscr{I}$ (Fig. 23). Any point P near $\mathscr{I}$ lies on timelike curves to the future and past which reach $\mathscr{I}$ in a region near to P (from the point of view of $\overline{M}$). Thus, any candidate $\mathscr{S}$ for a GCH would have to

---

[26] One proof of this is given by Seifert [98].

[27] This will be shown elsewhere.

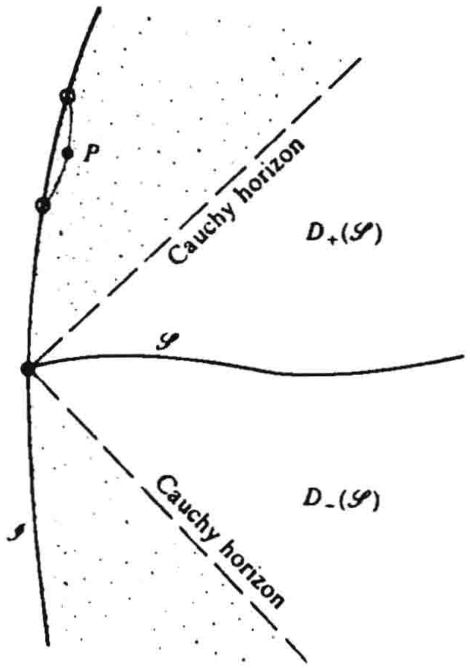


FIGURE 23. If $\mathscr{I}$ is timelike, every semispacelike set has a Cauchy horizon.

come near to P. Since this would apply to *all* points near $\mathscr{I}$, some of which have a definite timelike separation from each other, we see that this cannot be achieved with a semispacelike $\mathscr{S}$.

The de Sitter and anti-de Sitter space-times provide explicit examples where $\mathscr{I}$ is spacelike and timelike, respectively. To describe these models, we consider hyperspheres in five-dimensional pseudo-Euclidean space with two choices of signature. *De Sitter* space-time (Fig. 24) is given by

$$C^2 = -V^2 + W^2 + X^2 + Y^2 + Z^2 \qquad (9.14)$$

where the metric is

$$ds^2 = dV^2 - dW^2 - dX^2 - dY^2 - dZ^2 \qquad (9.15)$$

Anti-de Sitter space-time (Fig. 25) is the universal covering manifold of

$$C^2 = V^2 + W^2 - X^2 - Y^2 - Z^2 \qquad (9.16)$$

with metric

$$ds^2 = dV^2 + dW^2 - dX^2 - dY^2 - dZ^2 \qquad (9.17)$$

De Sitter space-time is topologically $S^3 \times E^1$ and possesses S^3 sections which are GCH's. But every timelike curve has both a particle horizon

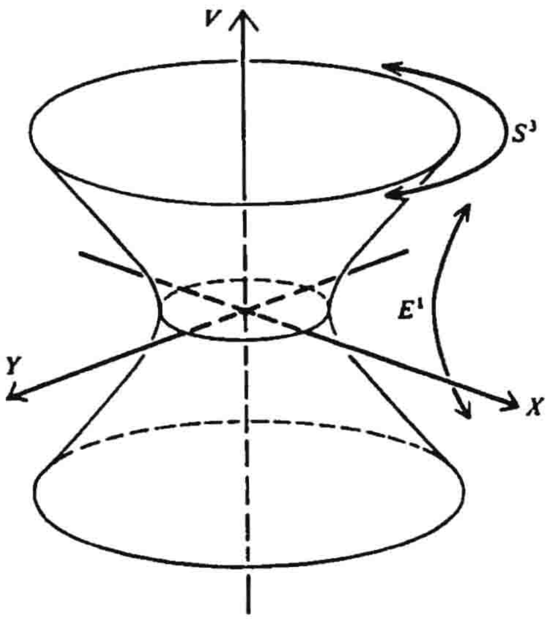


FIGURE 24. The de Sitter model.

and an event horizon. We can represent the space-time conformally on the Einstein static universe as the region between two parallel spacelike sections (Fig. 26). This evidently gives both $\mathscr{I}^+$ and $\mathscr{I}^-$ as *spacelike*. The steady-state model is just the "top-half" of the de Sitter model (see [97], for example), where we cut it at the intersection with, say, the hyperplane $V = W$ in (9.14) (Fig. 27). This cut leaves a *null* boundary in the past, so some timelike curves (for example, the substratum particles) do *not* possess particle horizons.

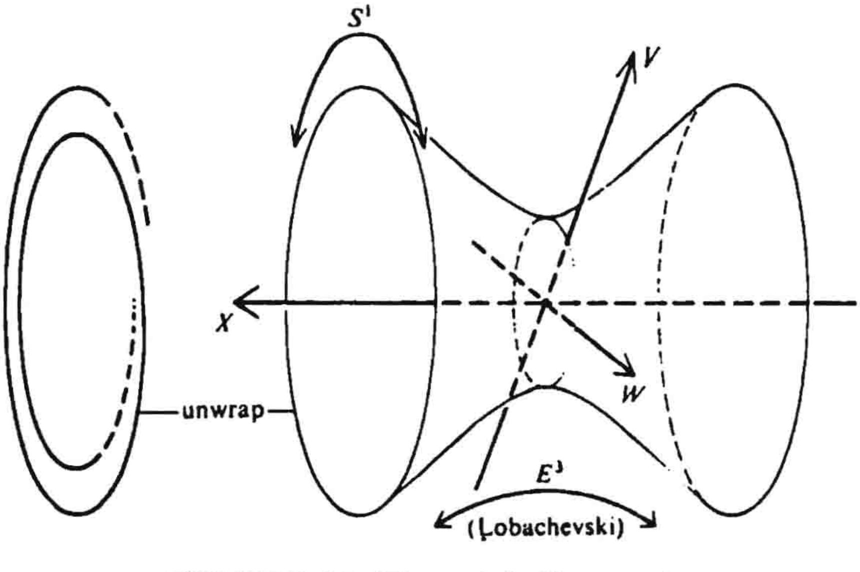


FIGURE 25. The anti-de Sitter model.

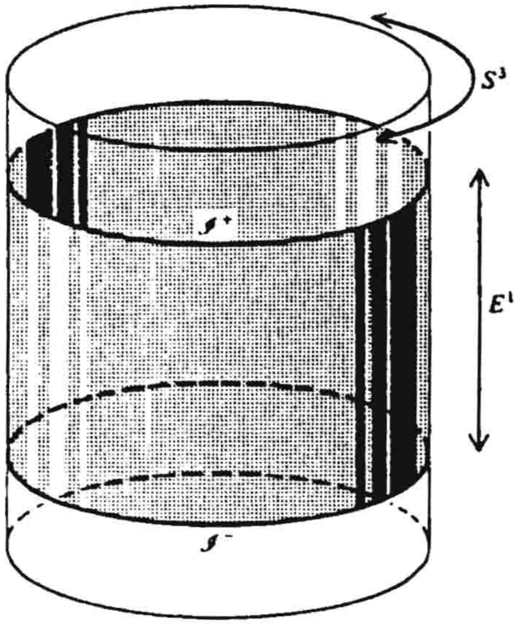


FIGURE 26. The de Sitter model as a conformal portion of the Einstein universe. The Friedmann model with $\lambda = 0$, $k > 0$ is similarly represented.

The sections of (9.14) by $V - W =$ const. > 0 give the (flat) hypersurfaces of constant time of the steady-state model. Unlike the entire de Sitter and anti-de Sitter space-times, this model is not geodesically complete.

The hypersphere (9.16) has topology $S^1 \times E^3$ and is thus not simply-connected. It possesses closed timelike curves (S^1) which can be opened out by taking the universal covering manifold. Then, when represented

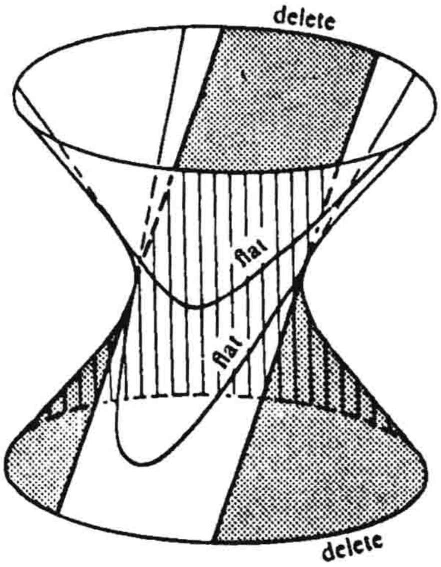


FIGURE 27. The steady-state model.

conformally on the Einstein static universe $S^3 \times E^1$ the "unwrapped" anti-de Sitter model becomes just a hemisphere of the S^3 times the E^1 (Fig. 28). The topology is thus just E^4, but its *conformal* structure is globally different from that of Minkowski space-time. We have a *timelike* $\mathscr{I}$, so every semispacelike $\mathscr{S}$ has a Cauchy horizon. On the other hand, no inextendible timelike geodesics possess particle or event horizons (although some inextendible timelike curves do. The existence of Cauchy horizons is connected with an interesting feature of the anti-de Sitter model. Let P be any point on the hypersphere (9.16) and let Q be diametrically opposite to P. Any plane through PQ cuts the hypersphere in a "great circle" which therefore must be a geodesic. One can see without great difficulty that, in fact, every timelike geodesic on the hypersphere, through P, must also pass through Q, but spacelike or null geodesics through P go to infinity and do not reach Q. We may pass to the "unwrapped" anti-de Sitter space (the universal covering space). For simplicity, choose versions of P and Q as near as possible, with $P \ll Q$ (Fig. 29). Choose R to be a point separated from Q by a spacelike geodesic. Then it is easy to convince oneself that there is *no geodesic connecting P to R* [17]. (We can get from P to R in two steps,

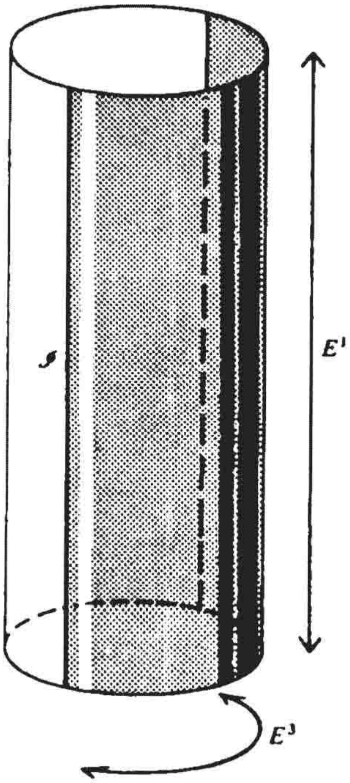


FIGURE 28. The anti-de Sitter model as a conformal portion of the Einstein universe.

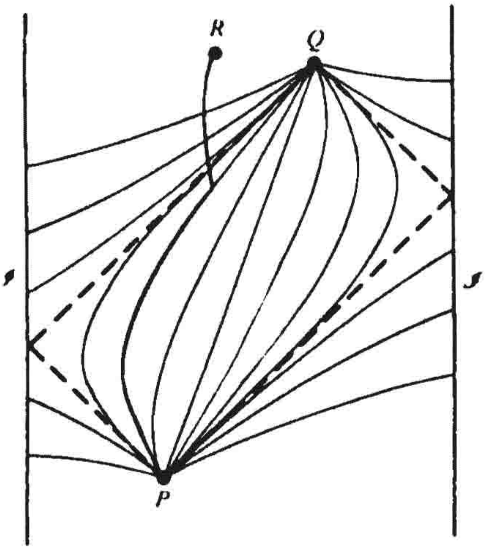


FIGURE 29. The points P and R cannot be joined by a geodesic.

however, using a broken timelike geodesic.) The anti-de Sitter model is, in fact, geodesically complete, so the situation is here very different from the case of positive-definite metrics. The nonexistence of a geodesic from P to R is closely related to the fact that R lies beyond $H_+(\dot{I}_+(P))$, as we shall see in Section 11.

The models most often studied in cosmological contexts are the Friedmann models, with metric

$$ds^2 = dt^2 - R^2\, d\Sigma^2 \tag{9.18}$$

Here $d\Sigma^2$ stands for the metric of a unit 3-sphere ($k = 1$), or of a Euclidean 3-space ($k = 0$), or a unit hyperbolic (Lobachevski) 3-space ($k = -1$). The quantity $R = R(t)$ satisfies

$$\frac{4}{3}\pi\rho R^3 = M = \text{const.} > 0 \tag{9.19}$$

$$\dot{R}^2 = \frac{KM}{2\pi R} + \frac{1}{3}\lambda R^2 - k \tag{9.20}$$

which gives an energy-momentum tensor appropriate to "dust" (that is, a perfect fluid without pressure):

$$T_{ab} = \rho(t) t_a t_b \tag{9.21}$$

where

$$t_a = \nabla_a t \tag{9.22}$$

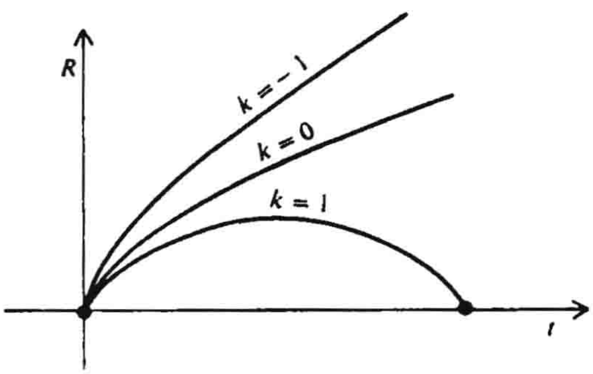


FIGURE 30. The different time-behaviors for the Friedmann models with $\lambda = 0$.

With $\lambda = 0$, the solutions of (9.20) are as indicated in Fig. 30. (The time-reverses of these are also solutions, but not consistent, if $k \leqslant 0$, with the observation that the universe is expanding.) Note that only in the case of a universe with positively curved (and therefore compact) space sections does the model collapse back to a singularity ($R = 0$). (By means of identifications, the cases $k \leqslant 0$ can also be given compact space sections, if desired, at the expense of a loss of global isotropy.)

We can represent the Friedmann models also as conformal subsets of the Einstein static universe. Let us take $\lambda = 0$. The case $k > 0$ resembles

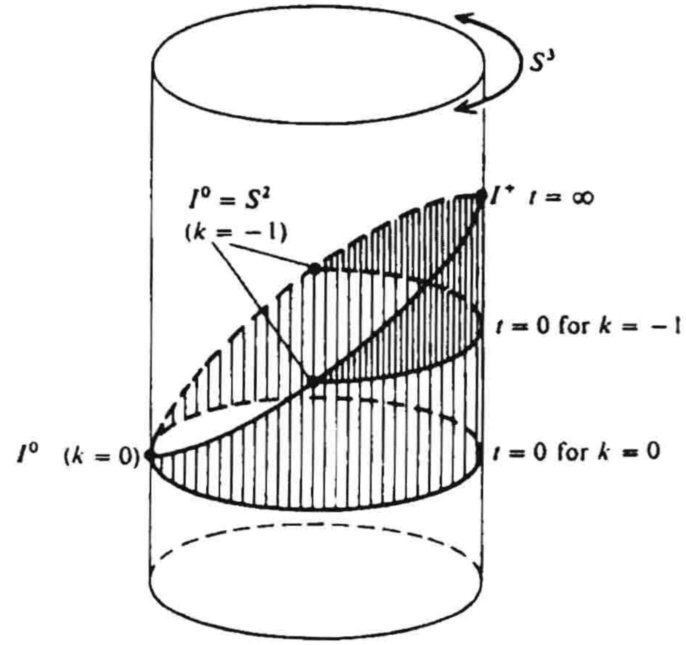


FIGURE 31. The Friedmann models with $k \leqslant 0$, $\lambda = 0$ as conformal portions of the Einstein universe.

the situation for de Sitter space-time (Fig. 26) the difference being that $\Omega = \infty$ at $\mathscr{I}^{\pm}$ instead of $\Omega = 0$. The cases $k \leqslant 0$ are indicated in Fig. 31. In both cases $\mathscr{I}^{-}$ is spacelike with $\Omega = \infty$ and $\mathscr{I}^{+}$ is null with $\Omega = 0$. When $k = 0$ we get $\overline{\nabla}_a \Omega = 0$ at $\mathscr{I}^{+}$, and we have a point I^0. But $\overline{\nabla}_a \Omega \neq 0$ when $k < 0$ and I^0 is an S^2. All the Friedmann models possess GCH's, but they yield particle horizons for all timelike curves.

Let us finally consider two examples which do not possess GCH's, but which fail to do so in rather different ways. First, we consider the plane-fronted wave, which has metric [15, 28, 77]

$$ds^2 = 2\{du + H(v, x, y)\,dv\}\,dv - dx^2 - dy^2 \tag{9.23}$$

the ranges of the real variables u, v, x, y being unrestricted. The nonvanishing curvature tensor components for (9.23) are defined by

$$\frac{\partial^2 H}{\partial x^2} \qquad \frac{\partial^2 H}{\partial x\,\partial y} \qquad \frac{\partial^2 H}{\partial y^2} \tag{9.24}$$

The metric satisfies Einstein's vacuum equations with $\lambda = 0$ (and so represents a purely gravitational wave) if

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) H = 0 \tag{9.25}$$

but if (9.25) is not imposed then (9.23) covers the more general situation of a combined gravitational–electromagnetic–neutrino wave. All seven principal null directions then coincide. If H is quadratic in x, y, then we get the special case of a *plane wave*. Many plane waves are geodesically complete [28, p. 96].

If $H = 0$ over some range, we have a region of Minkowski space-time. We may, in fact, choose H to be zero over all but some finite range of v: $v_0 < v < v_1$. Then we have a "sandwich wave" [10]. An extreme case is the idealized situation in which the wave is allowed to become infinitesimal in duration (say $v_0, v_1 \to 0$) while still producing a nonzero resultant effect. This results in the function H becoming a delta function in v:

$$H(v, x, y) = \delta(v) h(x, y) \tag{9.26}$$

With the substitution of (9.26) in (9.23), we get a metric which does not satisfy the conditions normally required for a space-time with a delta function in the *curvature*, as required in (9.24), since here the *metric* tensor components involve delta functions. It is possible to make a coordinate change so that the new metric tensor components are C^0 functions of the coordinates. But the present form is useful since it leads to a graphic "scissors and paste" construction of the manifold [80].

Divide Minkowski space-time, with metric

$$ds^2 = 2du\,dv - dx^2 - dy^2 \tag{9.27}$$

into $\mathscr{M}^-(v < 0)$ and $\mathscr{M}^+(v > 0)$ by removing the null hyperplane $v = 0$. Reattach a boundary $v = 0$ to each half but identify the two halves so that each appears "warped" as viewed from the other (Fig. 32). This is achieved, on passing from $\mathscr{M}^-$ to $\mathscr{M}^+$, by

$$x \to x \qquad y \to y \qquad u \to u + h(x, y) \tag{9.28}$$

at the common boundary $v = 0$. The jump in coordinates given in (9.28) corresponds precisely to (9.26) substituted in (9.23).

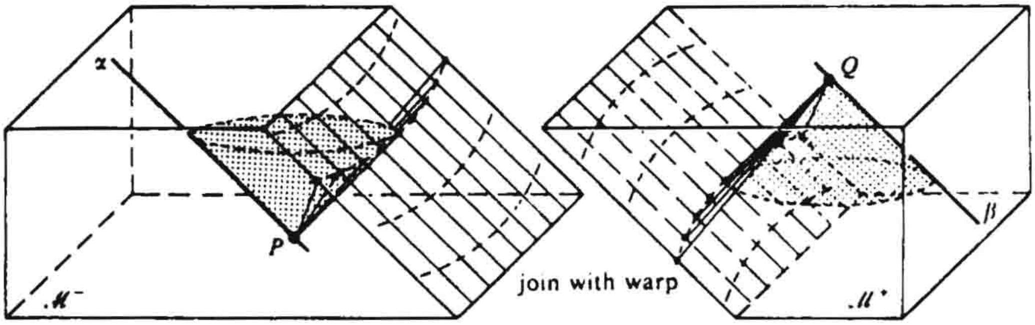


FIGURE 32. A plane wave with a delta-function amplitude. The two half-spaces are flat, but joined with a warp.

For simplicity, let us consider the purely electromagnetic plane wave case. Then we can take

$$h(x, y) = a(x^2 + y^2) \tag{9.29}$$

where $a =$ const. > 0. The null cone of the point $P \in \mathscr{M}^-$, with coordinates $u = x = y = 0$, $v = -a^{-1}$ has the equation

$$2u(v + a^{-1}) = x^2 + y^2 \tag{9.30}$$

At $v = 0$, this agrees with

$$2u(v - a^{-1}) = x^2 + y^2 \tag{9.31}$$

by virtue of (9.28) and (9.29). Equation (9.31) gives the null cone of the point $Q \in \mathscr{M}^+$ with coordinates $u = x = y = 0$, $v = a^{-1}$. This shows that the null cone of P is (in this example) exactly focused again, after passing through the wave, to the point Q. However, there is one exceptional null geodesic α, through P, namely $x = y = 0$, $v = -a^{-1}$, which never hits the wave at all. Correspondingly, the null geodesic β through Q, given by $x = y = 0$, $v = a^{-1}$ does not emerge on the other side. We have the situation that the *two* inextendible null geodesics α, β arise as the limiting configuration of a sequence of *single* null geodesics through both P and Q. (The space of null geodesics, in this example, is thus not Hausdorff.)

A spacelike GCH would have to intersect each null geodesic of the sequence just once. Thus it would have to meet the limit pair α, β in only *one* point. It could not meet both of α, β as would be required of a GCH. In terms of Fig. 32, any edgeless spacelike hypersurface through P gets trapped underneath the past null cone of Q and never reaches β. The nonexistence of a semispacelike GCH is also easily established. (This particular example can also be represented conformally as part of the Einstein static universe. In this case, the nonexistence of a GCH arises, not from a spacelike $\mathscr{I}$, but from the fact that $\dot{\mathscr{M}}$ contains a null geodesic in the *interior* region of $\bar{\mathscr{M}}$.) The more general conformally curved plane waves, such as the purely gravitational plane waves do not possess GCH's for essentially the same reason as the above case [77], but the situation becomes somewhat more complicated owing to the presence of astigmatism in the focusing.

As the final example of this section, we consider the Taub–NUT empty universe as described by Misner [59, 61, 63, 105] (see the lecture of Misner Ch. VI; Geroch Ch. VIII). Here, the entire space-time M is to be topologically $S^3 \times E^1$ (but conformally different from the Einstein universe). There is a particular spacelike section $\mathscr{S}$, of maximal proper volume, which has the metric of a 3-sphere (Fig. 33). The four-dimensional geometry at $\mathscr{S}$ does not possess the full rotational symmetry of S^3, however. The Weyl tensor is nonzero, the gravitational principal null directions coinciding in pairs {22} [*cf.* (7.6)]. The orthogonal projection into $\mathscr{S}$ of these null directions gives us a nowhere vanishing line-element field on $\mathscr{S}$. (The two null directions project as diametrically opposite, at each point of $\mathscr{S}$.) These line elements are tangential to a congruence of great circles on S^3 which form a set of *Clifford parallels*. Thus, these circles form a Hopf fibering of S^3. The rotational symmetry that remains is a transitive four-parameter group.

Let us move our section parallelly upward, in Fig. 33, preserving this symmetry. We get similar spacelike sections but where the volume of the S^3 decreases. This volume finally decreases to zero [in accordance with the Raychaudhuri effect, *cf.* (7.40)–(7.47)], but instead of collapsing to a point we simply get $H_+(\mathscr{S})$, which is, here, a compact null hypersurface, generated by null geodesics which constitute a set of Clifford parallels on $H_+(\mathscr{S}) \cong S^3$. Beyond $H_+(\mathscr{S})$, the section becomes timelike[28] and therefore possesses closed timelike curves, in a gross violation of reasonable causality.

---

[28] These sections are compact Lorentzian 3-manifolds of a type considered by Avez [3]. Any closed timelike curve γ of this 3-manifold possess a curious property, namely, that whereas (clearly) it cannot be continuously deformed to a point while remaining timelike, there is (for each section) some integer n with the property that $n\gamma$ (that is, γ described with multiplicity n— for this, γ needs to be a proper parametrized curve) can be deformed remaining timelike into $(n + 1)\gamma$. Hence, it can be so deformed into a curve of arbitrarily great length. This provides a counter-example to a result of Avez [4]. I am indebted to W. Kundt and H. J. Seifert for pointing this out to me.

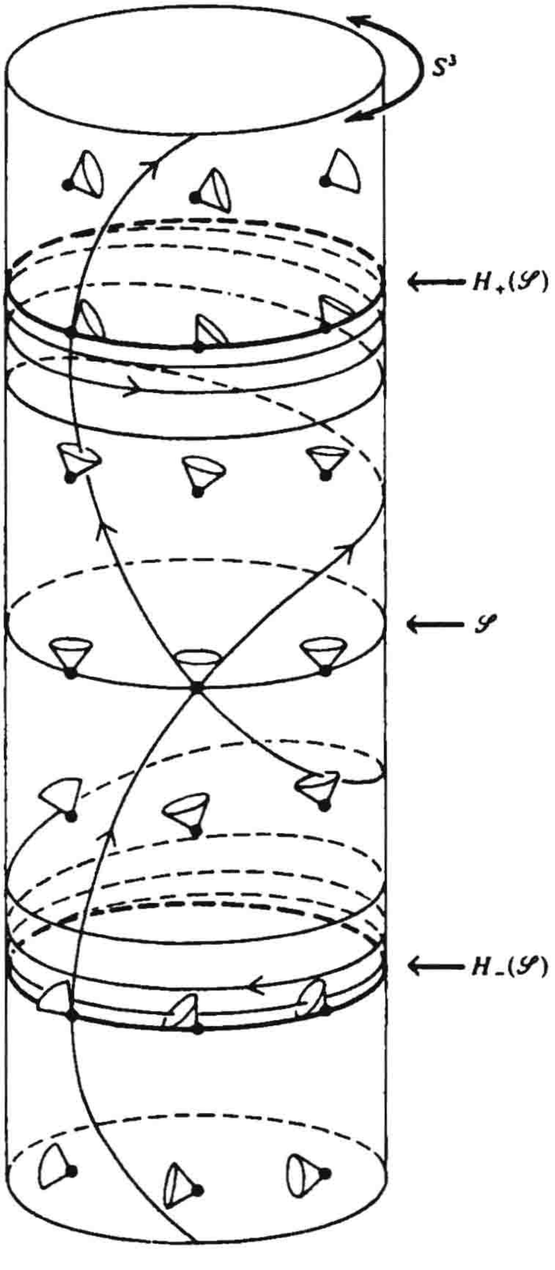


FIGURE 33. The Taub–NUT model.

The future-directed null and timelike geodesics emanating from points of $\mathscr{S}$ fall into two classes. There are those which enter the part of $\mathscr{M}$ beyond $H_+(\mathscr{S})$ (and so cross $H_+(\mathscr{S})$ at a definite point) and those which simply spiral round and round and approach $H_+(\mathscr{S})$ asymptotically. In fact, these latter curves have finite affine length into the future, so the model is not geodesically complete. Misner has shown that, curiously enough, an alternative extension can be fitted at $H_+(\mathscr{S})$ and beyond, other than the one shown in Fig. 33. The curves which previously crossed $H_+(\mathscr{S})$ now spiral round and round, while most of the geodesics which previuosly spiraled now cross the new $H_+(\mathscr{S})$ into the new extension. Both extensions are analytic, but can be added simultaneously only at the expense of giving up the requirement that $\mathscr{M}$ be a Hausdorff manifold [59, 61, Geroch lecture this volume, Ch. VIII]. The behavior at $H_-(\mathscr{S})$ is exactly similar to that at $H_+(\mathscr{S})$ (see Fig. 33). This model appears to be highly unstable, in the neighborhood of $H_\pm(\mathscr{S})$, against small perturbations in the initial conditions at $\mathscr{S}$. However, it is instructive in that it exhibits a kind of pathology that has to be taken into consideration when more general situations are studied.

10 GRAVITATIONAL COLLAPSE

A question of great interest in relativity theory and cosmology is that of the existence of singularities. The usual "big bang" models of relativistic cosmology are characterized by the presence of an initial state of infinite curvature and infinite density (*cf.* Fig. 30). At such an initial singularity the ordinary view of space-time as a smooth manifold would have to break down. Also, the existence of a particle horizon for each world line presents us with an essential problem. Different highly curved portions of the universe, which previously could not communicate with one another, must come together in such a way that they fit consistently. The only previous "causal" link between these portions was the initial singularity itself. We are presented with a problem of *principle* in addition to all the formidable, but essentially technical problems, concerned with the physics of early high-density states of these models. This is the question of *initial conditions* in an essentially singular situation. (See Prof. Misner's lecture, this volume, Ch. VI; also [60].)

It is a reasonable question to ask, however, whether the initial singularity of these models is not more a consequence of the mathematical idealizations involved, than of the actual physical situation that the model is supposed to describe. In particular, it is virtually essential to postulate a high degree of symmetry (for example, spatial homogeneity), for an exact model to be amenable to detailed mathematical treatment. And with such symmetry present, the picture is almost inescapable that all the matter in the universe should, at some time, have been compressed simultaneously into a single point, or at least into a line or into a two-dimensional surface [6, 43, 47a, 99]. Possible ways out of the dilemma that have been suggested include the adoption of a large enough cosmological constant, the presence of "unreasonable" matter (for example, negative energy densities—models with continual creation of matter throughout space would have to come under this heading), perhaps sufficient rotation, or a change in the laws of physics which could have been operative at very early stages of the universe. In addition, there are possible quantum effects of gravitation to be considered, although these might only be important at "ridiculously" high densities (for example, 10^{93} g/cc, or, at least, 10^{53} g/cc). But more significant in the first instance than any of these other considerations, is the role played by local irregularities of matter and curvature. In fact, near the singularity of the model, we really have no reason at all to expect that it should be realistic. The situation is perhaps clearer for the time-reversed process. If we imagine all the matter in the universe hurtling simultaneously to a single "central point" (or "line," etc.) then an infinite density at that point (or line, etc.) is hardly surprising. If, however, we perturb slightly the motion of the incoming matter, so that it is not exactly focused to one point (line, etc.) then we might expect that although the density could become very high, it would not become infinite. The possibility then may be considered of an effective "bounce" to the universe. Thus the presently observed expanding state of the universe might be regarded as the result of a previously contracting phase, the transition from contraction to expansion being achieved via a condensed but highly asymmetrical and complicated intermediate phase.

A picture of this kind has been suggested by many authors [54]. Also some detailed considerations by Lifshitz and Khalatnikov [53] had suggested that such avoidance of singularities by small perturbations of the motions might indeed be possible. However, recent work has ruled out most possibilities of this kind, as we shall see in Section 11. In the present section, I shall discuss the closely related phenomenon of the singularities encountered in gravitational collapse. The essential similarity between gravitational collapse and the time-reverse of the early stages of a "big bang" cosmology has been emphasized particularly by Wheeler [113]. Indeed, there are, perhaps, certain advantages in concentrating attention on the more local phenomenon of collapse, rather than on the global cosmological question. For we do not need to know the entire large-scale structure of the universe to study a local phenomenon (or do we?); if the local laws of physics are strongly affected by the structure of the universe as a whole ("Mach's principle"), then we might at least expect the local phenomenon to be the *less sensitive* to such changes in these laws; finally, unlike the universe as a whole, collapse is *not* something given to us just *once*, so we are freed to exploit the immense theoretical potentialities of the "gedanken experiment." My own personal feeling is that the deeper aspects of the cosmological problem will not be solved without a simultaneously deeper appreciation of what is involved in collapse.

But what is meant, precisely, by "gravitational collapse"? The idea comes originally from the classical study of large spherically symmetrical bodies. According to the work of Chandrasekhar [21, 39], a cold spherically symmetrical (nonrotating) body, which has reached the endpoint of thermonuclear activity, cannot hold itself apart against gravitational forces if its mass is much greater than that of the sun. Such a body would therefore collapse catastrophically and its collapse could only be halted if by some mechanism it were able to throw off sufficient mass to fall below Chandrasekhar's limit. In fact, certain detailed calculations (for references see [107]) have indicated that for initial masses between about 1.2 and 15 times that of the sun, the extreme intensity of the neutrino radiation from the collapsing core may be sufficient to blow off the outer regions of the star and leave a *neutron* star at the center. A cold neutron star (which is essentially an enormous atomic nucleus, bound together gravitationally rather than with nuclear forces) must actually be somewhat *less* massive than the sun (Oppenheimer–Volkoff limit [67]), though vastly smaller, with diameter of only about 10 kilometers. The gigantic explosion involved in such a process is thought to account for (at least) one type of supernova [107].

If no more matter falls into it, a neutron star could exist indefinitely. But what of larger initial masses? Stars with masses of the order of sixty times that of the sun are observed, for example. According to the above-mentioned calculations, the indications are that the core remains *too* massive to form a neutron star and therefore continues to collapse. Assuming that the *exact* spherical symmetry is maintained, the core would fall radially inward *through* its Schwarzschild radius $r = 2m$, to encounter a space-time singularity at the center $r = 0$. The remaining parts of the star which are not blown off would then follow the core in. An observer situated at large distances from the star would never see the collapse to within $r = 2m$, however. To him, the star's collapse would appear to slow down and approach $r = 2m$ asymptotically.

Before considering how physically realistic this picture is, let us examine the solution of Einstein's equations on which it is based, namely, the Schwarzschild solution. This defines the spherically symmetrical gravitational field outside the collapsing star. We have, with the usual Schwarzschild coordinates,

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2). \quad (10.1)$$

Let the star's boundary be given by $r = f(t)$. The metric (10.1) then only applies outside this boundary, that is, $r \geqslant f(t)$ (with the usual inequalities and identifications holding for θ and φ). The star's boundary must be timelike: $|f'(t)| < 1 - 2m/r$, from which it follows that this boundary never crosses $r = 2m$ in (10.1) (Fig. 34). One might be inclined to infer that the star therefore always remained outside $r = 2m$. But if we calculate the total proper time of a particle on the surface of the star, where we assume, for simplicity, that the star falls freely inward, we find that this proper time is *finite*. Thus, an observer who follows the star inward must, after this finite

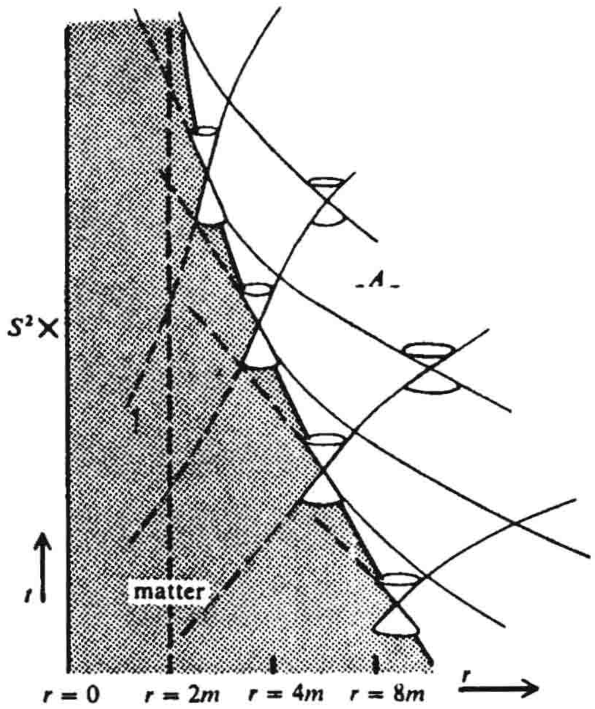


FIGURE 34. The Schwarzschild picture.

time has expired, either of necessity be destroyed (for example, because he encounters infinite tidal forces or some other form of singularity), or else, find himself in a portion of the universe not covered by the coordinates of (10.1).

In fact, in this instance it is the latter possibility which holds. We can see this if we transform the coordinates in (10.1) by introducing an advanced time parameter $v = t + r + 2m \log (r - 2m)$. Then the metric takes the Eddington–Finkelstein [27, 32] form [compare (8.14)]:

$$ds^2 = \left(1 - \frac{2m}{r}\right) dv^2 - 2\, dv\, dr - r^2(d\theta^2 + \sin^2 \theta\, d\varphi^2) \qquad (10.2)$$

When $r > 2m$, the forms (10.1) and (10.2) are precisely equivalent, but (10.2) has the advantage that the metric covers a larger region in a nonsingular way. The metric in the neighborhood of the (null) hypersurface $r = 2m$ is perfectly regular. It is only because the more familiar form (10.1) possesses a coordinate singularity at $r = 2m$, that this hypersurface is sometimes referred to as the "Schwarzschild singularity." The surface of the star crosses $r = 2m$ at a definite value of v (although, of course, the here unsuitable t parameter becomes infinite) and proceeds inward toward $r = 0$. But because of the positioning of the light cones (see Fig. 35), an outside observer cannot see any of the empty region $r \leqslant 2m$. More particularly,

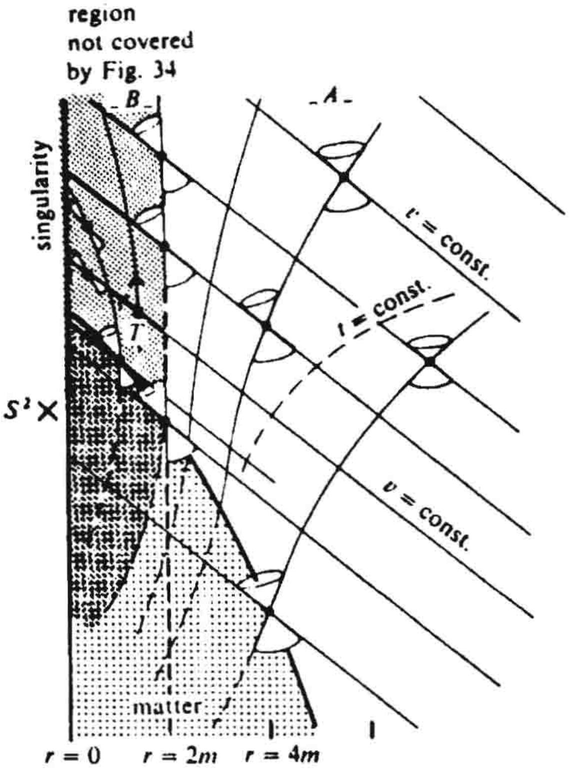


FIGURE 35. The Eddington–Finkelstein picture.

this is because, in the empty region, $r = 2m$ is a null hypersurface which lies "above" (that is, to the future of) the entire external region. An observer or signal can cross the hypersurface from the outside to the inside but not the other way around. The hypersurface $r = 2m$ is an *event horizon* of a more absolute character than most of those considered in Section 9. In fact, *all* geodesics which escape to infinity have this hypersurface as their event horizon.

If we trace the history of the star's surface inside the $r = 2m$ region, we see that it must, of necessity, encounter $r = 0$, for the surface must continue to move in a timelike direction and the null-cones are tipping over more and more toward the $r = 0$ axis. Again if we calculate the total proper time of a particle on the surface of the star (assuming free fall or otherwise), up until the point at which it encounters $r = 0$, we find that this is finite. But this time there is no hope of extending the solution further. If we evaluate the curvature scalars constructed from the Weyl tensor, we find that they tend to infinity as r approaches zero. Thus, our observer who successfully followed the star through $r = 2m$ must now be torn to pieces by the infinite tidal forces at the *true* singularity at $r = 0$.

This is the picture presented by Oppenheimer and Synder [68] when they considered the dynamics of a collapsing uniform dust cloud. In fact the space-time *interior* to such a dust cloud can also be described simply and explicitly and turns out to be nothing other than (a portion of) a Friedmann universe (see Section 9). This again emphasizes the close relationship between the situation of a final singularity in gravitational collapse and of a final (or initial) singularity for a relativistic cosmology.

If we wish to consider the entire history of a gravitationally bound Oppenheimer–Synder gas cloud and not just the final collapse phase, then even the metric form (10.2) does not cover a sufficient portion of the empty region. The behavior of such a dust cloud is time-symmetrical: Like the Friedmann universe with $k = 1$ ($\lambda = 0$), it starts from an initial singularity, expands to a maximum volume and then recontracts to a final singularity. Thus we may expect to join the matter-filled region to an empty, spherically symmetrical exterior region which is also time-symmetrical. We can, in fact, exhibit such a region if we use the U, V coordinates of Kruskal [50a], which are related to the Schwarzschild r, t (in the region $r > 2m$) by

$$\frac{V}{U} = -e^{t/2m} \qquad VU = e^{r/2m}\left(1 - \frac{r}{2m}\right) \tag{10.3}$$

The metric then takes the form

$$ds^2 = f^2\, dU\, dV - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \tag{10.4}$$

where

$$f^2 = \left(\frac{32m^3}{r}\right)e^{-r/2m} \tag{10.5}$$

The matter region is joined to the Kruskal region along a timelike hypersurface. This hypersurface is described by a timelike geodesic in the Kruskal (U, V)-plane (Fig. 36). Thus, only the portion to the right of this geodesic in the diagram is to be considered in this model. The left-hand portion of Fig. 36 must be replaced by a spherically symmetric portion ($\cong D^3 \times E^1$) of the Friedmann universe (9.18) whose boundary is again a hypersurface ($\cong S^2 \times E^1$) generated by timelike geodesics. These two portions do, in fact, fit together adequately smoothly, *cf.* [107].

One drawback of the Kruskal form of the Schwarzschild metric is that (10.4) is not given "explicitly," depending on the solution, for r, of (10.3). The Eddington–Finkelstein form (10.2) is, however, much simpler, and it is interesting that the main features of the Kruskal extension can be obtained simply from (10.2). To achieve this, we note that (10.2) covers the two portions A, B of the Kruskal diagram (Fig. 36) (with $2mV^2 = e^{v/2m}$). By using a retarded parameter u [*cf.* (8.11)] in place of the advanced parameter v, we can likewise cover the portion A, D of Fig. 36 with another "coordinate patch," the overlap region being A. In an exactly similar way, we can cover the portion C, B or the portion C, D with a "coordinate patch" having the form of (10.2). The coordinate transformations in the overlap regions C and

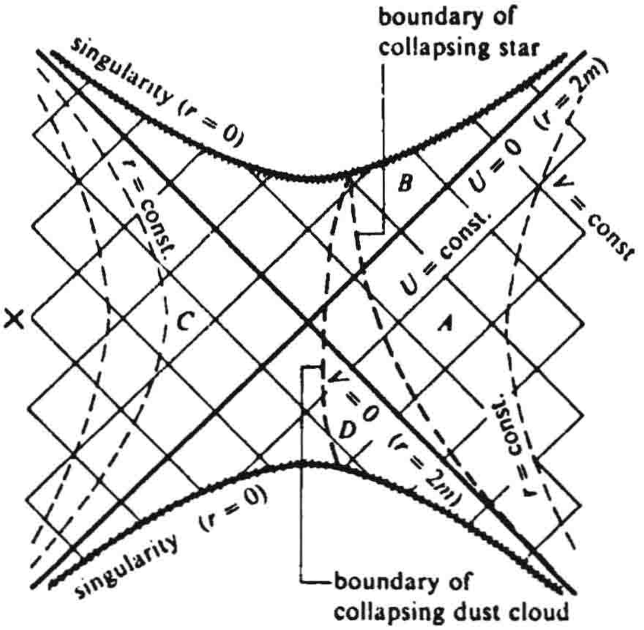

boundary of collapsing star
singularity (r = 0)
(r = 2m)
U = 0
B
r = const.
U = const.
V = const
S²×
C
A
V = 0
D
(r = 2m)
r = const.
singularity (r = 0)
boundary of collapsing dust cloud

FIGURE 36. The Kruskal picture.

D can be obtained, if desired, by transforming back to the form (10.1) (with $0 < r < 2m$) as an intermediate stage. The one feature of the Kruskal extension which is not obtained by this method is the regular nature of the space-time at the 2-sphere $U = V = 0$. However, the method has a considerable heuristic value in that it applies at once to certain other more complicated extensions, notably those of the Kerr and Reissner–Nordström solutions [*cf.* (10.18), (10.20)] [12, 18, 37]. In this connection it is useful to have also the version (Fig. 37) of the Kruskal diagram in which conformal

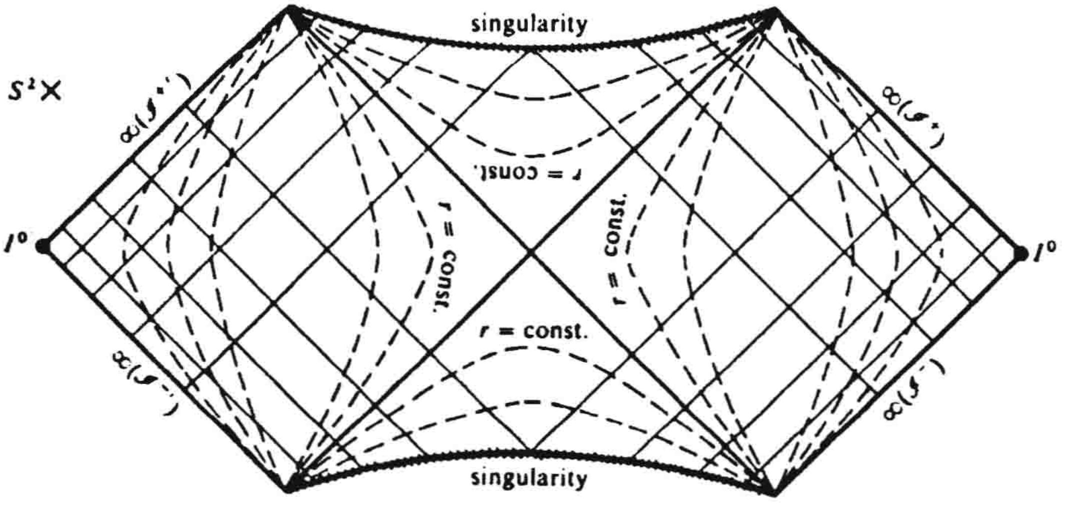

singularity
S²×
∞(𝒥⁺)
∞(𝒥⁺)
r = const.
r = const.
r = const.
r = const.
I⁰
I⁰
∞(𝒥⁻)
∞(𝒥⁻)
singularity

FIGURE 37. The Kruskal picture with conformal infinity represented.

infinity is represented (for example, use coordinates $p = \tan^{-1}\sinh^{-1} V$, $q = \tan^{-1}\sinh^{-1} U$).

Let us return, now, to the case of the collapsing star. Can the description that I have given be regarded as at all realistic? For example, even if we retain the exact spherical symmetry, can it be said that we really know enough about the properties of matter at the density (somewhat greater than that of an atomic neucleus, for the cases considered) at which the star is supposed to fall through its "Schwarzschild throat"? And might it not be that for some reason, a massive star inevitably ejects sufficient material, as it nears $r = 2m$, that its mass invariably falls below the Chandrasekhar or Oppenheimer–Volkoff limit, thus enabling a stable self-supporting state to become possible? It seems that for many years, the prevailing belief had been that a collapsing star would always be able to save itself by such means. This may have been partly because astronomers did not feel the need to consider self-supporting masses that were enormously greater than that of the sun. But with the discovery of the quasi-stellar objects (see, for example, [90]), renewed interest in the subject of gravitational collapse has been stimulated. For, considering the fantastic quantities of energy that these objects appear to be emitting, and their remarkably small size, it has been suggested that "individual" masses of the order, say, 10^6 or 10^8 times that of the sun may be involved. Now the Schwarzschild radius of a spherically symmetric body is proportional to its mass. Thus the characteristic density at which such a body crosses $r = 2m$ should be inversely proportional to the square of its mass. For an object of 10^6 or 10^8 solar masses, this density is not unreasonably high. Indeed, Fowler has emphasized that for an object of 10^{11} solar masses (such as a good-sized galaxy) this characteristic density is less than that of air! And there seems no reason at all to believe that such an object would necessarily eject virtually all its material before reaching $r = 2m$.

Then what about the role played by deviations from spherical symmetry? For example, can one even *define* the analog of $r = 2m$ for an asymmetrical body? And might it not be that the presence of *rotation* always prevents the final collapse? The question of rotation is a particularly pertinent one. For example, if differential rotation is present, then equilibrium states exist for objects of very much larger mass than Chandrasekhar's limit [69]. However, with viscosity present, only uniform rotation could represent a possible final state. Stability of such bodies is a crucial consideration [21a], and it appears that under suitable circumstances a rotating body can collapse. Now, fortunately, there is a known solution of Einstein's vacuum equations, namely that of Kerr [12, 48], which generalizes the Schwarzschild solution by the inclusion of *angular momentum*. The Kerr solution contains two arbitrary parameters m and a, where m defines the mass as in Schwarzschild's solution, and where ma defines the angular momentum. The solution [given explicitly in (10.18)] is still of a rather special character (for example, the quadrupole moment has to be given by $2ma^2$) but it seems likely that at least the main features of the geometry of the space-time surrounding a rotating collapsing body can be gleaned from a study of Kerr's solution. One of the most significant features of the solution, in fact, is that if $a > m$ there is no analog of the "Schwarzschild throat," that is, there is no event horizon for external observers, which would prevent signals from the interior regions being received outside. On the other hand, if $a < m$, there *is* such an event horizon and the solution qualitatively resembles the Schwarzschild solution in that it possesses a similar "throat" which can finally swallow up the material of the (super-) star. There are other features of the solution which qualitatively differ, however, from the case of the Schwarzschild solution. Some of these will be discussed shortly, but in the meantime, I shall concentrate on the question of the "throat" itself and give a definition which may be used as a criterion for its existence.

Let us return to the Schwarzschild solution. For convenience, consider the metric in the form (10.2) and examine Fig. 35. We wish to single out a property of the region B which characterizes it as "perculiar" in some sense. It should be emphasized, however, that there is nothing *locally* peculiar about the B-region. A small neighborhood of any point in the B-region is as good a solution of Einstein's vacuum equations as any other, and there is nothing at all "singular" about it. But it seems that in some way the whole B-region is "shrinking" in an apparently inevitable way to the space-time singularity at $r = 0$. To express this "shrinkage" we need to single out a partly *global* aspect of the B-region—but it must not be *too* global because we require a property that will not be destroyed by small perturbations in the metric at the "time" of collapse. Consider, then, a point T in the B-region of Fig. 35 which can lie just inside $r = 2m$ (but outside the matter region). Such a point describes a spacelike 2-sphere $\mathscr{T}$ in the whole space-time, of total surface area $4\pi r^2 (r < 2m)$. Now any system of material particles whose world-lines intersect $\mathscr{T}$ must have a subsequent motion which is bounded by the velocity of light. That is to say, if we consider the extreme situation of a flash of (idealized) light emitted at $\mathscr{T}$, there will be an "outgoing" flash and an "ingoing" flash, described by the two null geodesic segments in Fig. 35 with past endpoint $\mathscr{T}$. These two null geodesic segments describe null hypersurfaces through $\mathscr{T}$, which together constitute $\dot{I}_+[\mathscr{T}]$. Now the significant feature is that *both* parts of the boundary are "shrinking." Not only is the total surface area of the sphere described by a point moving up the "ingoing" null geodesic decreasing, but so also is the total surface area correspondingly decreasing for the "outgoing" flash. This, then, is the property of $\mathscr{T}$ which I shall use to characterize it as a *trapped surface* [75]. The presence of a trapped surface in a region will then be our indication that there may be something "peculiar" about the subsequent history of the region.

CHAPTER 36

To be more precise:

A TRAPPED SURFACE $\mathscr{T}$ IS A SMOOTH COMPACT, SPACELIKE 2-SURFACE[29] WITH THE PROPERTY THAT THE NULL GEODESICS, WHICH MEET T ORTHOGONALLY LOCALLY CONVERGE IN FUTURE DIRECTIONS. (10.6)

The quantity ρ of Section 7 conveniently measures thic convergence. The null geodesics meeting a spacelike 2-surface orthogonally always generate a null hypersurface, so by (7.28) we have $\rho = \bar{\rho}$. Thus, the characteristic feature of a trapped surface is that

$$\rho > 0 \quad \text{AT} \quad \mathscr{T} \tag{10.7}$$

for both the null hypersurfaces through $\mathscr{T}$. As we move into the future up either of these null hypersurfaces, the surface area decreases, initially, at every point of $\mathscr{T}$.

The idea, here, is to establish a connection between the existence of a trapped surface in a space-time and the subsequent development either of a singularity in space-time [such as $r = 0$ in (10.2)] which may involve infinite curvature, or else of some other space-time behavior which would be very strange from the physical point of view. But before considering the rigorous mathematical result, let us ask the question whether it is reasonable to expect trapped surfaces to develop at all in our actual universe. For any *individual* case of a collapsing star or superstar, this is clearly a matter of the detailed astrophysics involved. The amount of rotation present, the mass loss via neutrinos, through gravitational radiation, the exact nature of the asymmetry present, the nature of the equations of state, magnetic fields, etc.; all these may be relevant. But the essential point I want to make here is that there can be no reason *of principle* against a trapped surface developing. To illustrate this, I shall consider the following "*gedanken experiment.*"

A technically advanced (but presumably exceedingly foolhardy) race of beings inhabit a galaxy (preferably elliptical) which contains about 10^{11} stars. By means of rockets, these beings contrive to alter the velocities of the stars, so that there remains practically no transverse component to the stars' motions. Furthermore, their radial motions are adjusted so that all the stars will fall toward the center so as to reach its vicinity at almost the same time. (The mass-energy expended in this operation would be small compared to the mass of the stars.) The region they have to aim for has roughly a radius 50 times that of the solar system. There is plenty of room in a volume of this size for all the stars, and they can arrive there before any serious problem of stellar collisions arises. (They can even steer to avoid collisions, if necessary!) But does this result in the development of a trapped surface? The fact that it does depends on the most primitive considerations of general relativity. We consider a flash of (idealized) light emitted at the center at about the time the stars enter the critical region. Because of the focusing effect on idealized light rays [7, 87] (that is, on null geodesics) by the gravitational field of each star (which is actually an *observed* effect of general relativity), a sufficient density of light rays which started off by diverging from the center will pass through (and near to) enough stars so that they start converging again. Under these circumstances it can be shown that a topological sphere which lies just inside the outermost boundary of the light rays will indeed be a trapped surface.

---

[29] It is convenient also to assume that $\mathscr{T}$ is a semispacelike set, but this is not strictly necessary for the proof of Theorem I.

This example serves to illustrate, also, the fact that there is no reason why an observer should be "destroyed" as he enters the critical region. The curvature is still extremely small in the neighborhood of the trapped surface and the space-time there is perfectly regular.[30] Indeed, locally there is nothing peculiar about the trapped surface itself. Even in Minkowski space-time, many surfaces exist which are "locally trapped." That is to say, they satisfy the conditions of (10.6) *except* for the fact that they are not compact. To construct such a surface, we need only consider the intersection of the past null cones of two spatially separated points.

If, then, it is accepted that trapped surfaces can exist, what of the consequences? I shall show, in fact, that if we assume (provisionally) that the universe is *open* (that is, not spatially compact), then our present knowledge of the laws of physics is insufficient to enable us to calculate even *in principle* the future history of such a collapsing system.

To predict the future of a system, we would normally ask that adequate Cauchy data be specified on some spacelike hypersurface. We may take it that the relevant local laws of physics are such that effects can be propagated along timelike or null curves, but not along spacelike ones. Thus, if data are specified on some semispacelike closed set $\mathscr{S}$, then this can determine the situation only in $D_+(\mathscr{S})$ and $D_-(\mathscr{S})$. For regions exterior to these, we could have relevant information carried along nonspacelike curves which fails to "register" on $\mathscr{S}$. Now a collapse situation can be set up starting from perfectly reasonable data on a spacelike hypersurface $\mathscr{C}$, which (since we are trying to consider a local physical situation, rather than a cosmological one) we attempt to think of as *effectively* extending to infinity whether or not the universe is actually spatially noncompact. Thus, for *predictability*, we ask that a noncompact spacelike hypersurface $\mathscr{C}$ should exist which is a GCH for $\mathscr{M}$, or, more appropriately, for which $D_+(\mathscr{C}) = I_+[\mathscr{C}]$. The question of whether the cosmologial nature of space-time is *actually* relevant to a local collapse situation is not a trivial one, as we shall see later.

---

[30] In [26], slightly aspherical collapse is examined and it is verified that trapped surfaces evolve.

Theorem I. The following requirements on a space-time[31] $\mathscr{M}$ are mutually inconsistent:

(Ia) There exists a noncompact spacelike hypersurface $\mathscr{C}$ for which $D_+(\mathscr{C}) = I_+[\mathscr{C}]$
(Ib) There exists a trapped surface $\mathscr{T} \subset I_+[\mathscr{C}]$
(Ic) $R_{ab}l^a l^b \leqslant 0$ for each null vector l^a
(Id) $\mathscr{M}$ is future-null complete

In (Id), "future-null complete" means that every null geodesic in $\mathscr{M}$ can be extended into the future to arbitrarily large values of a (given) affine parameter. We may think of this physically as the statement that "photons (or neutrinos or gravitons) cannot just disappear." This seems to be a very reasonable physical requirement on a space-time. In Section 11, we have to consider another type of completeness, namely, timelike geodesic completeness ("particles in inertial motion cannot just disappear"). Yet other types are possible, such as "bounded acceleration completeness" which states that all inextendible timelike curves of bounded curvature have infinite total length[32] (that is, "particles experiencing limited forces cannot just disappear"). These types of completeness are all *inequivalent* for suitably contrived examples [35, 52]. However, it probably does not make much physical difference which definition is used. It is possible that the definitions are only inequivalent in the presence of regions of arbitrarily large curvature. To establish the existence of such regions would really be nearer to the object of results such as Theorems I, II, III, in any case. Unfortunately no result which *directly* predicts the existence of such high-curvature regions, on the basis of "reasonable" and "generic" physical assumptions, has yet been established, although we may think of Theorem I, and particularly Theorems II and III, as strong indirect indications of this; *cf.* Geroch lecture Ch. VIII.

The condition $R_{ab}l^a l^b \leqslant 0$ is simply (7.35), and is a consequence of Einstein's equations (7.1) (with or without the λ-term) and the positive definiteness of the energy density (7.34). (In an "eigentetrad" of T_{ab}, this positive definiteness means that $T_{00} \geqslant 0$, $T_{00} + T_{11} \geqslant 0$, $T_{00} + T_{22} \geqslant 0$, $T_{00} + T_{33} \geqslant 0$.) This is very desirable physically, especially since a violation of this inequality seems to lead to a serious problem in connection with quantum field theory, namely, an apparent catastrophic instability of the vacuum. But the issue is not completely clear-cut.

---

[31] Space-times considered here are always assumed to be time-orientable (*cf.* Section 5). However, all the results here (for example, Theorems I, II, III) will also apply in a suitable form without this assumption. For any non-time-orientable Lorentzian manifold, there always exists a double covering manifold which is time-orientable. (Consider the space of all null half-cones on the manifold. This gives the desired double covering, *cf.* [55].)

[32] The term "length" will often be used here to denote "proper time" for a timelike curve.

Before we proceed to the proof of Theorem I, it will be useful to establish some lemmas concerning the structure of a SSB. We recall [*cf.* (9.4)–(9.6)] that any SSB is a subset $\mathscr{S}$ of $\mathscr{M}$ of the form

$$\mathscr{S} = \dot{I}_+[\mathscr{K}] \text{ or, equivalently, } \mathscr{S} = \dot{I}_-[\mathscr{L}] \tag{10.8}$$

where $\mathscr{K}, \mathscr{L} \subset \mathscr{M}$. Let us divide $\mathscr{S}$ into four disjoint parts (some of which may be vacuous), $\mathscr{S}_N, \mathscr{S}_+, \mathscr{S}_-$, and $\mathscr{S}_0$, defined as follows. If $X \in \mathscr{S}$, then there *may* or *may not* exist points $Y, Z \in \mathscr{S}$, distinct from X, for which

$$I_+(X) \subset I_+(Y) \qquad I_-(X) \subset I_-(Z) \tag{10.9}$$

the different possibilities defining the subsets $\mathscr{S}_N$, $\mathscr{S}_+$, $\mathscr{S}_-$, $\mathscr{S}_0$ according to the scheme

$$X \in \begin{array}{c|c|c|c} & (\exists Z) & (\nexists Z) & \\ \hline & \mathscr{S}_N & \mathscr{S}_+ & (\exists Y) \\ \hline & \mathscr{S}_- & \mathscr{S}_0 & (\nexists Y) \end{array} \tag{10.10}$$

The intuitive meanings of these subsets is that $\mathscr{S}_N$ represents the portion of $\mathscr{S}$ which is *null*, where $\mathscr{S}_+$ and $\mathscr{S}_-$ consist, respectively, of the future and past end-points of $\mathscr{S}_N$, and that $\mathscr{S}_0$ represents the spacelike portion of $\mathscr{S}$ (see Fig. 38). A more precise statement is contained in the following lemma,

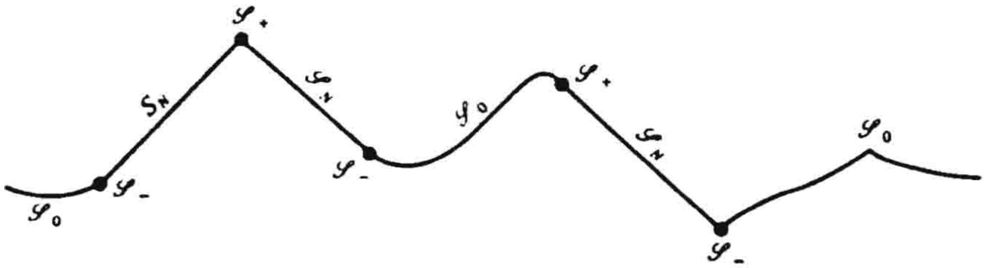


FIGURE 38. The different portions of a semispacelike boundary. (Null directions are inclined at 45°.)

which also gives a somewhat more useful condition for a point to lie in $\mathscr{S}_N$, $\mathscr{S}_+$, or $\mathscr{S}_-$.

Lemma I. Let $X \in \mathscr{S}$, where $\mathscr{S}$ is an SSB given as in (10.8). Let $\mathscr{Q}$ be an open set containing X. Then

(a) $I_+(X) \subset I_+[\mathscr{K} - \mathscr{Q}]$ implies $X \in \mathscr{S}_N \cup \mathscr{S}_+$
(b) $I_-(X) \subset I_-[\mathscr{L} - \mathscr{Q}]$ implies $X \in \mathscr{S}_N \cup \mathscr{S}_-$

Furthermore, a null geodesic segment on $\mathscr{S}$

(c) passes through X if $X \in \mathscr{S}_N$
(d) has X as future end-point if $X \in \mathscr{S}_+$
(e) has X as past end-point if $X \in \mathscr{S}_-$

To prove Lemma I, assume, first, that $I_+(X) \subset I_+[\mathscr{K} - \mathscr{Q}]$. Let $\mathscr{B}$ be a small normal coordinate ball with center X and boundary $\dot{\mathscr{B}}$, where $\bar{\mathscr{B}} \subset \mathscr{Q}$. Consider a sequence (X_n) of points, with $X_n \in I_+(X) \cap \mathscr{B}$ which converges to X. For each X_n, there exists $W_n \in \mathscr{K} - \mathscr{Q}$ with $W_n \ll X_n$. The timelike curve from W_n to X_n contains a connected portion in $\mathscr{B}$, with past end-point $Y_n \in \dot{\mathscr{B}}$ and future end-point X_n. Now $\dot{\mathscr{B}}$ is compact, so there is a point Y on $\dot{\mathscr{B}}$ such that a subsequence of (Y_n) converges to Y. The join of Y, X in the normal coordinate system is a geodesic k, through X. It is readily seen that k cannot be spacelike (for small $\mathscr{B}$). Also k cannot be timelike, since then we should have $Y \ll X$, so $W_n \ll Y_n \ll X$ for some n, violating $X \notin I_+[\mathscr{K}]$. Thus, k is a null geodesic. Hence $I_+(X) \subset I_+(Y)$. Now, $Y_n \in I_+[\mathscr{K}]$, so $Y \in \bar{I}_+[\mathscr{K}]$. Also $Y \notin I_+[\mathscr{K}]$, since $X \notin I_+[\mathscr{K}]$, so that $Y \in \mathscr{S}$. This establishes (a), (d) and half of (c). The proof of (d), (e) and the other half of (c) is exactly similar. (That the two geodesics thus obtained for (c) must coincide, follows since $\mathscr{S}$ is semispacelike.)

When *two* null geodesic segments on $\mathscr{S}$ have a point X in common, this can only happen at $X \in \mathscr{S}_+$ or $X \in \mathscr{S}_-$. This result also holds in the case of "infinitesimally neighboring" null geodesic segments on $\mathscr{S}$. The following result makes this more precise.

Lemma II. Let $\mathscr{S}$ be an SSB and k a null geodesic segment lying on $\mathscr{S}$, with a future [resp. past] end-point P. Suppose some open set $\mathscr{Q}$, containing $k - P$, meets $\mathscr{S}$ in a smooth null hypersurface $\mathscr{N}$ for which the convergence ρ is unbounded near P (choosing the tangent vector l^a to k smoothly at P). Then $P \in \mathscr{S}_+$ [resp. $\mathscr{S}_-$].

The proof of Lemma II that I give depends on another result (actually more detailed than is necessary):

Lemma III. Let k be a null geodesic segment in $\mathscr{M}$ and let $\mathscr{N}_1$, $\mathscr{N}_2$ be smooth null hypersurfaces through k for which $\mathscr{N}_1 \subset I_+[\mathscr{N}_2] \cup \mathscr{N}_2$. Then $\rho_2 - \rho_1 \geqslant |\sigma_2 - \sigma_1|$ on k (ρ_i, σ_i refer to $\mathscr{N}_i$, with the same l^a for each $\mathscr{N}_i$ at k; $i = 1, 2$).

To prove Lemma III, consider two smooth scalar functions $u_i (i = 1, 2)$ on $\mathscr{M}$, where $u_i = 0$ defines $\mathscr{N}_i$ (in some neighborhood of k), with u_i increasing toward the future with nonvanishing gradient on $\mathscr{N}_i$. We can scale the u_i's so that $\nabla_a u_1 = \nabla_a u_2 = l_a$ on k. In a small enough neighborhood of k we have $u_2 \geqslant u_1$. It follows that $X^a \nabla_a (u_2 - u_1) + \frac{1}{2} X^a X^b \nabla_a \nabla_b (u_2 - u_1) \geqslant 0$, on k, for each set of components (X^a) with $|X^a|$ small enough. Let $\nabla_a u_i = l_{(i)a}$, so $l_{(1)a} = l_{(2)a} = l_a$ on k. Choose m_a, on k, complex, null and orthogonal to l_a, with $m_a \bar{m}^a = -1$. Putting $X^a = \lambda m^a + \bar{\lambda} \bar{m}^a$, we get $\lambda^2(\sigma_2 - \sigma_1) + 2\lambda\bar{\lambda}(\rho_2 - \rho_1) + \bar{\lambda}^2(\bar{\sigma}_2 - \bar{\sigma}_1) \geqslant 0$ [*cf.* (7.33)]. This must hold independently of λ, and $\rho_2 - \rho_1 \geqslant |\sigma_2 - \sigma_1|$ follows.

To prove Lemma II, suppose $P \in \mathscr{S}_N$ (with P a *future* end-point of k). Then, k can be extended into the future on $\mathscr{S}$ to a point $R \in \mathscr{S}$. The past null-cone $\dot{I}_-(R)$ is nonsingular in some open set $\mathscr{G}$ with $k - P \subset \mathscr{G} \subset \mathscr{Q}$ (for R near enough to P). Lemma III now applies with $\mathscr{N}_2 = \mathscr{G} \cap \dot{I}_-(R)$ and $\mathscr{N}_1 = \mathscr{G} \cap \mathscr{N}$. But if $\rho = \rho_1$ is unbounded near P (and positive, since Φ is not unbounded at P), then $\rho_2 - \rho_1$ must become negative near P, violating the inequality in Lemma III. Thus Lemma II is established.

One immediate consequence of Lemma I is that if $\mathscr{T}$ is any closed subset of $\mathscr{M}$ and

$$\mathscr{S} = \dot{I}_+[\mathscr{T}] \tag{10.11}$$

then,

$$\mathscr{S}_- \cup \mathscr{S}_0 \subset \mathscr{T} \tag{10.12}$$

This follows from part (a) since if $X \in \mathscr{S} - \mathscr{T}$, some neighborhood $\mathscr{Q}$ of X will not meet $\mathscr{T}$, so $\mathscr{K} - \mathscr{Q} = \mathscr{K}$, with $\mathscr{K} = \mathscr{T}$. From (10.12) and parts (c) and (d), it follows that each point of $\mathscr{S} - \mathscr{T}$ is the *future end-point of a null geodesic segment on* $\mathscr{S}$. As particular cases, this result applies to $\dot{I}_+(P)$ or to any particle horizon. We may also take the time-reverse of the result, so if $\mathscr{S} = \dot{I}_-[\mathscr{T}]$, then each point of $\mathscr{S} - \mathscr{T}$ is the past end-point of a null geodesic segment on $\mathscr{S}$. This applies, in particular, to $\dot{I}_-(P)$ or to any event horizon.

Another consequence of Lemma I concerns the structure of Cauchy horizons. The result is not needed for Theorem I, but will be required in the next section. There are two simple ways of obtaining any Cauchy horizon as a part of an SSB. Let $\mathscr{H}$ be closed and semispacelike. Then the set $\operatorname{int}\{D_-(\mathscr{H}) \cup I_+[\mathscr{H}]\}$ is the union of the $I_+(P)$ with $P \in D_-[\mathscr{H}]$. The boundary of this set is an SSB which contains the Cauchy horizon $H_-(\mathscr{H})$ as part (that is, as the intersection with $D_-(\mathscr{H})$). This is the first way. The second way can be described as follows. Let $\mathscr{H}$ be closed and semispacelike, as before, and define [42]

$$W_-(\mathscr{H}) = I_-(\mathscr{H}) - D_-(\mathscr{H}) \tag{10.13}$$

(We can define $W_+(\mathscr{H})$ correspondingly.) Then $P \in W_-(\mathscr{H})$ if and only if P is both the past end-point of a future-inextendible timelike curve α *not* meeting $\mathscr{H}$, and also the past end-point of a timelike curve β with future end-point *on* $\mathscr{H}$ (Fig. 39). Clearly $W_-(\mathscr{H})$ is open. Also, if $Q \ll P$ and $P \in W_-(\mathscr{H})$, then $Q \in W_-(\mathscr{H})$, since $\mathscr{H}$ is semispacelike. Hence,

$$W_-(\mathscr{H}) = I_-[W_-(\mathscr{H})] \tag{10.14}$$

Thus $\dot{W}_-(\mathscr{H})$ is an SSB. It is readily verified that

$$H_-(\mathscr{H}) = \dot{W}_-(\mathscr{H}) \cap D_-(\mathscr{H}) \tag{10.15}$$

[use (9.5)], so again, our Cauchy horizon is obtained as a part of an SSB.

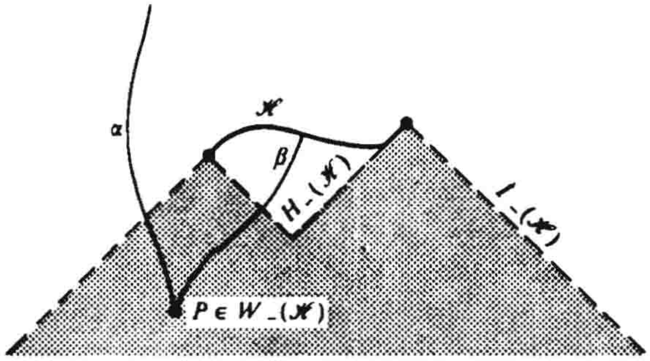


FIGURE 39. A semispacelike boundary containing the past Cauchy horizon of $\mathscr{H}$.

Now suppose $X \in W_-(\mathscr{H})$, but $X \notin \text{edge}\,(\mathscr{H})$ [*cf.* (9.13), etc.] Either an α-curve or a β-curve with past end-point X must exist (trivially—where we define a degenerate β-curve to exist whenever $X \in \mathscr{H}$) but not both, since $X \notin W_-(\mathscr{H})$. If it is the α-curve that fails to exist, then $X \in H_-(\mathscr{H})$; if it is the β-curve, then $X \in \dot{I}_-[\mathscr{H}] - \mathscr{H}$. Suppose $X \in H_-(\mathscr{H})$, so a β-curve (possibly degenerate) exists from X. Then we can choose a small open neighborhood $\mathscr{D}$ of X so that every point of $\bar{\mathscr{D}} \cap \dot{I}_-[\mathscr{H}]$ is the end-point of a β-curve (that is, $\dot{I}_-[\mathscr{H}] \cap \bar{\mathscr{D}} = \mathscr{H} \cap \bar{\mathscr{D}}$). For if $X \in \mathscr{H}$, then, since $X \notin$ edge $(\mathscr{H})$, there exist points A, B near X (with $A \ll X \ll B$) so that every timelike curve from A to B, in the neighborhood of X, meets $\mathscr{H}$. In this case we choose $\bar{\mathscr{D}} \subset I_-(B) \cap I_+(A)$. If $X \notin \mathscr{H}$, then choose $\bar{\mathscr{D}}$ within the past of the β-curve from X. Any point U of $\mathscr{D} \cap I_-(X)$ is in $W_-(\mathscr{H})$, so U is the past end-point of an α-curve, which must meet $\mathscr{D} \cap \dot{I}_-[\mathscr{H}]$ in V, say. Now V is the past end-point of a β-curve also, so $V \in \dot{\mathscr{D}} \cap W_-(\mathscr{H}) \subset W_-(\mathscr{H}) - \mathscr{D}$. Thus Lemma I condition (b) is satisfied, whence:

EVERY POINT OF $H_-(\mathscr{H})$ – edge $(\mathscr{H})$ IS THE PAST END-POINT OF A NULL GEODESIC SEGMENT ON $H_-(\mathscr{H})$ (10.16)

Of course, the same applies also to $\dot{I}_-[\mathscr{H}] - \mathscr{H}$. Note that if $H_-(\mathscr{H})$ does not meet edge $(\mathscr{H})$, then every point of $H_-(\mathscr{H})$ is the past end-point of a *future-inextendible* null geodesic on $H_-(\mathscr{H})$ (since extending the null geodesic into the future, we can never reach a future end-point on $H_-(\mathscr{H})$). This applies [*cf.* (9.13)] when $\mathscr{H}$ is an *edgeless semispacelike closed set.* The results will clearly also apply in the time-reversed versions.

Another consequence of (10.16) is that if $\mathscr{H}$ *is a semispacelike set which intersects every maximal null geodesic of* $\mathscr{M}$ *in a nonvacuous compact set, then* $\mathscr{H}$ *is a GCH for* $\mathscr{M}$. For, if $\mathscr{H}$ were not a GCH, it would possess a nonvacuous Cauchy horizon, a null geodesic generator of which would have to meet $\mathscr{H}$ in a compact set. An end-point of this set would have to lie on edge$(\mathscr{H})$, by (10.16). However edge$(\mathscr{H})$ must be vacuous. To see this, consider $\dot{I}_-[\mathscr{H}]$. Now $\mathscr{H} \subset \dot{I}_-[\mathscr{H}]$ since $\mathscr{H}$ is semispacelike. Suppose $P \in \dot{I}_-[\mathscr{H}]$ but $P \notin \mathscr{H}$. A null geodesic generator γ, of $\dot{I}_-[\mathscr{H}]$, passes through P and contains a separate compact portion of $\mathscr{H}$. Let Q lie on γ between P and this portion. Any null geodesic η through Q, distinct from γ can meet $\dot{I}_-[\mathscr{H}]$ only at Q. Thus η could not meet $\mathscr{H}$, so P does *not* exist. We have $\mathscr{H} = \dot{I}_-[\mathscr{H}]$, so $\mathscr{H}$ is an SSB; $\mathscr{H}$ is therefore edgeless and the result is established.

As a converse, we note that if $\mathscr{H}$ is any semispacelike set and $P \in \operatorname{int} D_-(\mathscr{H})$ then any future-inextendible null geodesic γ through P must meet $\mathscr{H}$. This follows because a point $Q \in D_-(\mathscr{H})$ exists with $Q \ll P$, so if γ did not meet $\mathscr{H}$ we could follow just below γ with a future-inextendible *timelike* curve, contradicting $Q \in D_-(\mathscr{H})$. Similarly, we have the time reverse of this result that any past-inextendible null geodesic through any point of int $D_+(\mathscr{H})$ must also meet $\mathscr{H}$. This will be required in a moment.

PROOF OF THEOREM I. Set $\mathscr{S} = \dot{I}_+[\mathscr{T}]$ as in (10.11). Then if $P \in \mathscr{S} - \mathscr{T}$, there will be a null geodesic k, a segment of which lies on $\mathscr{S}$ and has P as a future end-point. Extend the geodesic maximally into the past. It hits $\mathscr{C}$ because $P \in \operatorname{int} D_+(\mathscr{C})$. Since $\mathscr{S}(\subset I_+[\mathscr{C}])$ is closed, there is a point $R \in \mathscr{S}$ which is a past end-point of $k \cap \mathscr{S}$. This must lie on $\mathscr{S}_-$, so $R \in \mathscr{T}$. Furthermore k must meet the spacelike 2-surface $\mathscr{T}$ orthogonally at R, that is, lie on one of the two null hypersurfaces which intersect locally in $\mathscr{T}$, since these represent the local boundary of $I_+[\mathscr{T}]$. Thus, $\mathscr{S}$ is generated by the null geodesic segments $\hat{k}$ which are orthogonal to $\mathscr{T}$ at a past end-point, and which may have a future end-point (where they encounter $\mathscr{S}_+$). In fact, any such segment $\hat{k}$ *must* have a future end-point by Lemma II, because the trapped surface condition (10.7), together with the focusing property [*cf.* (7.39)] and null completeness (Id), imply that ρ becomes *unbounded at some first point* Q, on $\hat{k}$, or on the extension k of $\hat{k}$ into the future. (In fact, the second possibility is the more usual one, corresponding to the occurrence of a *crossing region* of $\mathscr{S}$, which generally occurs before the *caustic* is reached, where $\rho = \infty$.) Now Q varies continuously as k varies [*cf.* (7.29), (7.33), etc.] so the finite segments RQ of k generate a compact set. Thus $\mathscr{S}$, being a closed subset of a compact set, is also compact. So, we have (9.7) that:

$$\mathscr{S} \text{ IS A SEMISPACELIKE COMPACT TOPOLOGICAL 3-MANIFOLD IN } \mathscr{M} \qquad (10.17)$$

Now, it is a well-known theorem (see, for example, [101, p. 201]) that any space-time admits a smooth timelike unit vector field. We can use the integral curves of this vector field to achieve a one-to-one mapping of $\mathscr{S}$ into $\mathscr{C}$ (both sets are semispacelike) since, by (Ia), $\mathscr{S} \subset D_+(\mathscr{C})$. This is impossible since the dimensionalities are the same but $\mathscr{S}$ is compact and $\mathscr{C}$ is noncompact. Thus (Ia), (Ib), (Ic), (Id) are mutually inconsistent and Theorem I is established.

In the face of Theorem I, we are led to ask which of the conditions is most likely to break down in the actual universe. The most immediate candidate is perhaps (Ia). There are two ways that (Ia) can fail to be satisfied. In the first instance, the universe might be "closed" (that is to say, spatially compact). Then we might envisage replacing ($\mathscr{C}$) by a *compact* spacelike hypersurface which is a GCH (or with $I_+[\mathscr{C}] = D_+(\mathscr{C})$). The proof of Theorem I would then break down at the final stage. Indeed, the result would no longer be true, because (for example) all the conditions would be satisfied by the de Sitter space (9.14), (9.15), with $\mathscr{C}$ as the section of (9.14) by $V = -2a < 0$ (a const.) and $\mathscr{T}$ as $V = -a$, $W = 0$. On the other hand, it seems hard to believe that the question of whether the universe, as a whole, is open or closed should seriously affect the discussion of a "local" collapsing system (which would be smaller by a factor of at least 10^{11}). Indeed, an examination of the proof of Theorem I shows that the argument can still be made to carry through in the case of a compact $\mathscr{C}$ *unless*, in a well-defined sense, the collapsing object eventually "swallows up" the entire universe! However, a closed universe (like the Friedmann model with $k = 1$) which eventually recontracts to a highly condensed state could, in principle, satisfy this condition. But the elimination of the singularity problem by this means seems to be ruled out by a theorem of Hawking [41], which states, in effect, that "almost all" space-times with a *compact* GCH must be incomplete if they satisfy the slightly stronger (but still "reasonable") condition on R_{ab} implied by (7.47).

The second way in which (Ia) might be violated is a more serious possibility. The "Laplacian" idea that the future of the universe should be completely determined by its behavior at one "time" predates both relativity and quantum theory. We have become accustomed to nondeterminism in the latter theory, so why should we still demand determinism in all cases for general relativity? The possibility that the universe should not possess a GCH is an intriguing one. There are a number of physically interesting space-times which do not possess GCH's. Two examples were considered in Section 9. Other examples which are more relevant to the question of collapse (although, possessing singularities) are the Kerr [48] and Reissner–Nordström (see, for example, [62]) space-times, as extended, respectively, by Boyer and Lindquist [12] (*cf.* also [18, 19]) and by Graves and Brill [37]. The Kerr metric is

$$ds^2 = dv^2 - \frac{2mr}{h}(dv + a\sin^2\theta\, d\varphi)^2 - 2\, dv\, dr + h d\theta^2 + 2a\sin^2\theta\, dr\, d\varphi + (r^2 + a^2)\sin^2\theta\, d\varphi^2 \qquad (10.18)$$

where m and a are constant (mass m, angular momentum ma) and where

$$h = r^2 + a^2\cos^2\theta \qquad (10.19)$$

The Reissner–Nordström metric (with advanced time v and electric charge defined by e) is

$$ds^2 = \left(1 - \frac{2m}{r} + \frac{e^2}{r^2}\right) dv^2 - 2\,dv\,dr - r^2(d\theta^2 + \sin^2\theta\, d\varphi^2) \qquad (10.20)$$

The metric (10.18) with $m > |a|$ and the metric (10.20) with $m > |e|$ are somewhat similar in that they possess not only an event horizon for external geodesics (Schwarzschild "throat"), given, respectively, by

$$r_+ = m + (m^2 - a^2)^{1/2} \qquad r_+ = m + (m^2 - e^2)^{1/2} \qquad (10.21)$$

but also a second "horizon" at which the null cones tip back again:

$$r_- = m - (m^2 - a^2)^{1/2} \qquad r_- = m - (m^2 - e^2)^{1/2} \qquad (10.22)$$

Both metrics have true singularities at $r = 0$, but these differ somewhat in nature. In the case of the Kerr metric the singularity has a ring structure.

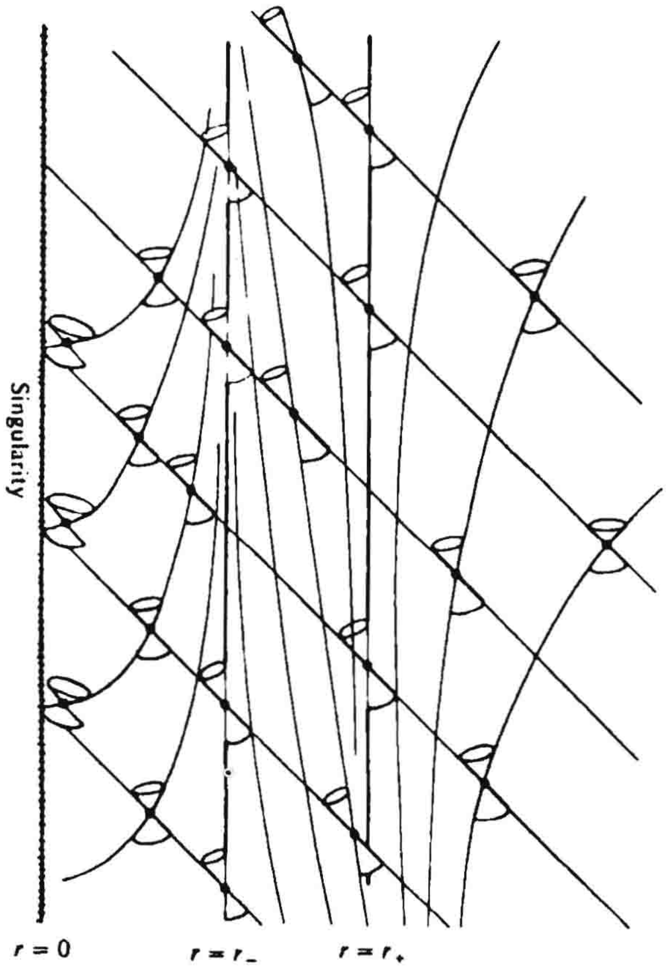


FIGURE 40. The Kerr picture (applicable also to the Reissner–Nordström metric).

By passing through the ring one enters a region with $r < 0$. (Then, as Carter [19] has shown, closed timelike curves are encountered near the ring.) Ignoring the nature of the metrics near (and through) the singularities, the situation is as in Fig. 40. By patchwork similar to that in the Schwarzschild case, but here more complicated, a " maximal " analytic extension is obtained (Fig. 41). The shaded portion in Fig. 41 represents the whole of Fig. 40. The entire Fig. 41 diagram can be pieced together (except for some exceptional central points) using overlapping patches isometric with the shaded region.

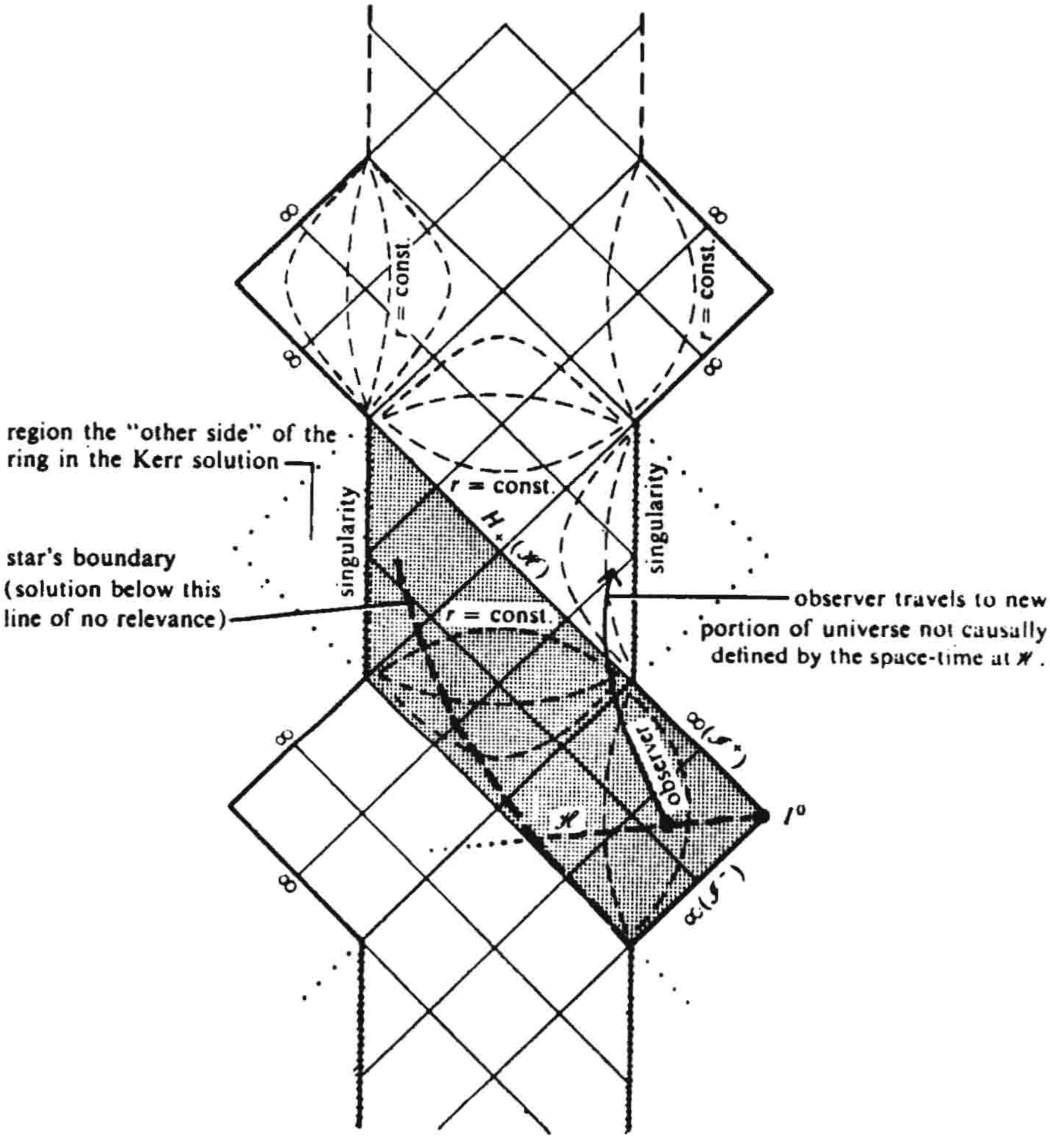

∞
r = const.
r = const.
∞
∞
∞
region the "other side" of the ring in the Kerr solution
star's boundary (solution below this line of no relevance)
singularity
r = const.
$H_+(\mathscr{H})$
r = const.
singularity
observer travels to new portion of universe not causally defined by the space-time at $\mathscr{H}$.
∞
observer
$\infty(\mathscr{I}^+)$
$\mathscr{H}$
I^0
∞
$\infty(\mathscr{I}^-)$

FIGURE 41. The Carter–Boyer–Lindquist picture (incorporating that of Graves–Brill).

(For this, one needs to be aware of a discrete symmetry of the metrics, which is not manifest in the forms (10.18), (10.20), *cf.* [12].)

Now, the significance of all this is that by following the star inward and then giving himself an acceleration outward after crossing $r = r_+$, an observer can find himself in a portion of the universe not causally determined by information on an apparently reasonable initial hypersurface $\mathscr{H}$. In fact, he crosses $H_+(\mathscr{H})$. The region he enters actually contains a singularity which he can "see." But of course, since he has crossed $H_+(\mathscr{H})$, we cannot guarantee that the space-time he enters is in fact the one that we have constructed for him. In effect, "another universe" has joined on to the universe he started from. The particular continuation we have chosen was obtained here by an *analytic* continuation rather than by the use of field equations. (But even analytic extensions do not ensure uniqueness globally. A striking example occurs with the Taub–NUT space-time described in Section 9, but also in flat space-time analytic extension does not imply uniqueness. The difficulty is that one does not know, in general, when to identify events reached in different ways.) Even if we can avoid the singularity in the "new universe" there is a problem of principle here, because, like in the early stages of a "big bang" cosmology, causally disconnected regions have to come together in such a way that they "fit." A striking example of this occurs with the Reissner–Nordström solution, since the "new universe" has to possess new charged particles of just the right total charge. (This follows even independently of the symmetry.)

There is a further difficulty confronting our observer who tries to cross $H_+(\mathscr{H})$. As he looks out at the universe that he is "leaving behind," he sees, in one final flash, as he crosses $H_+(\mathscr{H})$, the entire later history of the rest of his "old universe." It is here, in fact, that cosmological questions might be very relevant to him. If, for example, an unlimited amount of matter eventually falls into the star then presumably he will be confronted with an infinite density of matter along "$H_+(\mathscr{H})$." Even if only a finite amount of matter falls in, it may not be possible, in generic situations to avoid a curvature singularity in place of $H_+(\mathscr{H})$. This is at present an open question. But it may be, that the place to look for curvature singularities is in this region rather than (or as well as?) at the "center." In this connection, an example due to Bardeen [5] (*cf.* [23a, 66a]) of a collapsing charged dust cloud is of interest. Here *no* singularity occurs within $D_+(\mathscr{H})$, although a charge singularity has to be created in the "new universe" beyond $H_+(\mathscr{H})$ (see Fig. 41). Whether any reality can be attributed to this situation in which a "new universe" is created out of a collapse is an intriguing question. We might view (10.17) as an indication that a trapped surface results in a spatially compact universe "budding" off the original one. However, all this rests on the supposition that $H_+(\mathscr{H})$ can survive a generic perturbation on $\mathscr{H}$, without curvature singularities arising. We should also bear in mind that in any case it seems unlikely that curvature singularities can be avoided, in view of Theorem III of the next section.

11 SINGULARITIES IN COSMOLOGY

This section is devoted to two important theorems due to Hawking [40, 42]. The arguments will be given here partly in outline. For further details, the reader is referred to Hawking's original papers. The aim of the theorems is to show that on the basis of present physical understanding, there are (or were) very peculiar regions of space-time, probably involving enormous curvatures, at which we would have to expect a local physical behavior very different from that to which we have become accustomed. The interest of Theorem II will lie, perhaps, mainly in the fact that no causality assumptions whatever are made. It applies only to a spatially compact universe satisfying a condition (an inequality) which might possibly apply to the actual universe, but which would be virtually impossible to verify observationally. Theorem III, on the other hand, contains an analogous condition which, according to Hawking and Ellis [44], *is* likely to be satisfied on the basis even of present-day astronomical observations. In time-reversed form, Theorem III will also apply to suitable situations of gravitational collapse (such as a mildly perturbed Oppenheimer–Snyder collapsing dust cloud (see Section 10). A causality assumption is required for Theorem III, but of a much less severe kind than the global Cauchy hypersurface condition (Ia) of Theorem I. In effect, closed (that is, $\cong S^1$) timelike curves are to be excluded, but we shall require a slightly strengthened version of this.

Let $P \in M$, then the following conditions on $\mathscr{M}$ with respect to P are equivalent [42, 50]:

ARBITRARILY SMALL NEIGHBORHOODS OF P EXIST WHICH EACH INTERSECT NO TIMELIKE CURVE IN A DISCONNECTED SET (11.1)

that is, roughly speaking, timelike curves from the neighborhood of P cannot leave and then return to the neighborhood of P;

IF $P \prec Q$ AND EVERY POINT OF $I_-(Q)$ CHRONOLOGICALLY PRECEDES EVERY POINT OF $I_+(P)$, THEN $P = Q$ (11.2)

(*cf.* Section 9); or

THE TIME-REVERSE OF (11.2) (11.3)

If (11.1), (11.2), or (11.3) holds, then we say that *strong causality holds at* P. If the condition holds for *all* $P \in \mathscr{M}$, then we say *strong causality holds for* $\mathscr{M}$.

The equivalence of (11.1)–(11.3) is not hard to establish and will be omitted here. An example, due to Hawking, for which the condition fails is indicated in Fig. 42. (With Minkowski coordinates and metric, we omit ($x^0 > 1$, $x^1 \leqslant -1$), ($x^0 < -1$, $x^1 \geqslant 1$), ($x^0 = -x^1 = 1$) and ($x^0 = -x^1 = -1$); then identify $(1, x^1, x^2, x^3)$ with $(-1, -x^1, x^2, -x^3)$, $x^1 < -1$; P is the origin.) If strong causality holds at P, we call a neighborhood $\mathcal{Q}$ of P *causally convex* if no timelike curve meets $\mathcal{Q}$ in a disconnected set [*cf.* (11.1)].

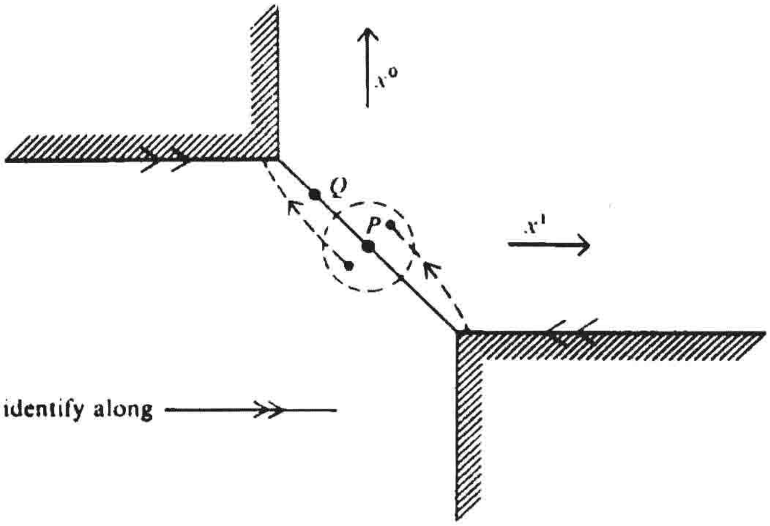


FIGURE 42. A future- and past-distinguishing space-time for which strong causality fails.

If strong causality fails at P, an arbitrarily small perturbation of the metric in the neighborhood of P could result in the actual presence of closed timelike curves, which for various reasons is highly undesirable physically. Thus, it appears to be a very reasonable supposition that strong causality should actually hold at every point of space-time. Another condition, which is weaker than (11.1)–(11.3) but stronger than nonexistence of a closed timelike curve through P is *future distinguishability* [50] at P:

$$I_+(P) = I_+(Q) \text{ IMPLIES } P = Q \qquad (11.4)$$

This states that timelike curves through P into the future, cannot leave and then return arbitrarily close to P. For the time-reverse of this, we say $\mathcal{M}$ is *past distinguishing at P* if and only if for all Q

$$I_-(P) = I_-(Q) \text{ IMPLIES } P = Q \qquad (11.5)$$

A hierarchy of causality conditions which are, on the other hand, stronger than (11.1), have been suggested by Carter.[33] Again, violation of any of Carter's conditions could lead to closed timelike curves when $\mathcal{M}$ is perturbed slightly.

The condition that strong causality should hold at every point of $\mathcal{M}$ is an interesting one. It is, in fact, equivalent to the statement that the "Alexandrov topology" $\mathcal{T}^*$ is Hausdorff [50], where the open sets of $\mathcal{T}^*$ have as a base, the sets $I_+(X) \cap I_-(Y)$. This is again equivalent to the statement that $\mathcal{T}^*$ agrees with the ordinary manifold topology of $\mathcal{M}$. We can refer to Fig. 42 for a situation under which the condition fails. For this example, any open set in $\mathcal{T}^*$, containing P, meets any open set, in $\mathcal{T}^*$ containing Q, so the Hansdorff property fails.

Let us now state the theorems.

Theorem II. (Hawking) The following requirements on a space-time $\mathcal{M}$ are mutually inconsistent:

(IIa) There exists a compact spacelike hypersurface (without boundary) $\mathcal{H}$.

(IIb) The divergence θ of the unit normals to $\mathcal{H}$ is positive at every point of $\mathcal{H}$.

(IIc) $R_{ab}t^a t^b \leqslant 0$ for each timelike vector t^a.

(IId) $\mathcal{M}$ is geodesically complete in past timelike directions.

We can think of (IIa) and (IIb) as stating that "the universe is (or was) spatially compact and expanding." However, the "expansion" as implied by (IIb) has to take place at *every* point of $\mathcal{H}$. Thus, in the presence of a collapsed object, we could not expect to be able to arrange (IIb) for an $\mathcal{H}$ "at the present time." If the condition is satisfied, therefore, $\mathcal{H}$ would probably have to refer to an early phase of the expansion.

On the assumption of Einstein's equations (7.1), requirement (IIc) is the "energy condition" (7.47), which (assuming, preferably, that $\lambda = 0$) is very reasonable physically.

Theorem III. (Hawking) The following requirements on a space-time $\mathcal{M}$ are mutually inconsistent:

(IIIa) Strong causality holds at every point of $\bar{I}_-(P)$, for some point $P \in M$.

(IIIb) The divergence of all the timelike and null geodesics through P changes sign somewhere to the past of P.

---

[33] One such condition would be violated if there were two points $P, R \in \mathcal{M}$, such that for any neighborhoods $\mathcal{P}$ of P and $\mathcal{R}$ of R, there would always be a timelike curve with past end-point in $\mathcal{P}$ and future end-point in $\mathcal{R}$, and another with future end-point in $\mathcal{P}$ and past end-point in $\mathcal{R}$. This behavior can occur without strong causality failing, as can be seen by taking $\mathcal{M}$ as the twofold covering manifold of that of Fig. 42 (Carter).

(IIIc) $R_{ab}t^a t^b \leqslant 0$ for each timelike vector t^a.

(IIId) $\mathcal{M}$ is geodesically complete in past timelike and null directions.

Condition (IIIb) is a little awkwardly stated because the divergence in the case of the timelike geodesics is measured by the quantity θ of (7.40) while the divergence of the null geodesics is measured by ρ of (7.18), (7.33). The region at which ρ changes sign is not actually the limit of the region where θ changes sign. Another way of stating (IIIb) would be to refer only to the timelike geodesics and state that the (relevant) change of sign takes place within a *compact region*.

I shall outline the proofs of Theorems II and III by indicating the main lemmas involved. (These lemmas and their proofs are all essentially due to Hawking [40, 42], although the discussion given here deviates slightly from his.)

Lemma IV. If a spacelike hypersurface $\mathscr{H}$ (without boundary) exists, which is a closed subset of $\mathcal{M}$ (that is, properly embedded), then a covering manifold $\mathcal{M}^*$ of $\mathcal{M}$ exists where the inverse image of $\mathscr{H}$ under the covering map consists of a set of discrete isometric copies of $\mathscr{H}$, each of which is semispacelike in $\mathcal{M}^*$.

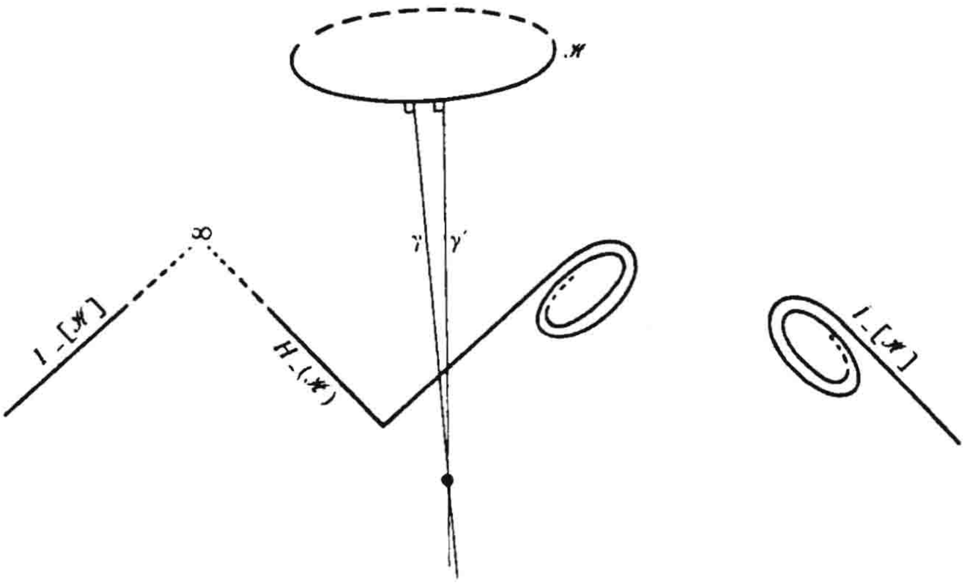
The idea is to make $\mathscr{H}$, in (IIa), into a semispacelike set, by "unwrapping" $\mathcal{M}$, but not so much as might render the new $\mathscr{H}$ noncompact. To achieve this, choose a point O in $\mathcal{M}$, then for each $P \in \mathcal{M}$ and for each equivalence class of curves from O to P whose intersection number with $\mathscr{H}$ is a given integer we assign a point of $\mathcal{M}^*$. (This intersection number is homotopy invariant because $\mathscr{H}$ has no edges.) One readily verifies that $\mathcal{M}^*$ has the properties required of it in the lemma. [This construction effectively obtains $\mathcal{M}^*$ from the universal covering manifold of $\mathcal{M}$, by putting an equivalence relation on the elements of the fundamental group of $\mathcal{M}$, saying two such elements are equivalent if they give equal intersection numbers with $\mathscr{H}$ (see also, Geroch [34]).]

Without loss of generality, we now regard $\mathcal{M}^*$ as $\mathcal{M}$ in Theorem II, so we can *take $\mathscr{H}$ to be a semispacelike set*. In order that we can treat Theorems II and III concurrently to a large extent, define, for Theorem III,

$$\mathscr{H} = \dot{I}_-(P) \tag{11.6}$$

(Strong causality at P implies that $\mathscr{H}$ is nonvacuous, resembling a null cone at P.) Then, in both cases, we have $\mathscr{H} = \bar{\mathscr{H}}$ as an edgeless semispacelike set.

It will be useful to have in the back of one's mind a picture of some types of situation which might possibly arise. For Theorem II, we envisage a situation like that of Fig. 43. Neither $H_-(\mathscr{H})$ nor $\dot{I}_-[\mathscr{H}]$ can meet $\mathscr{H}$ since $\mathscr{H}$ is edgeless and a spacelike hypersurface. Every point of $H_-(\mathscr{H})$ lies on a future-inextendible null geodesic on $H_-(\mathscr{H})$ which may go to infinity or (since causality assumptions are not made here) simply spiral round and round

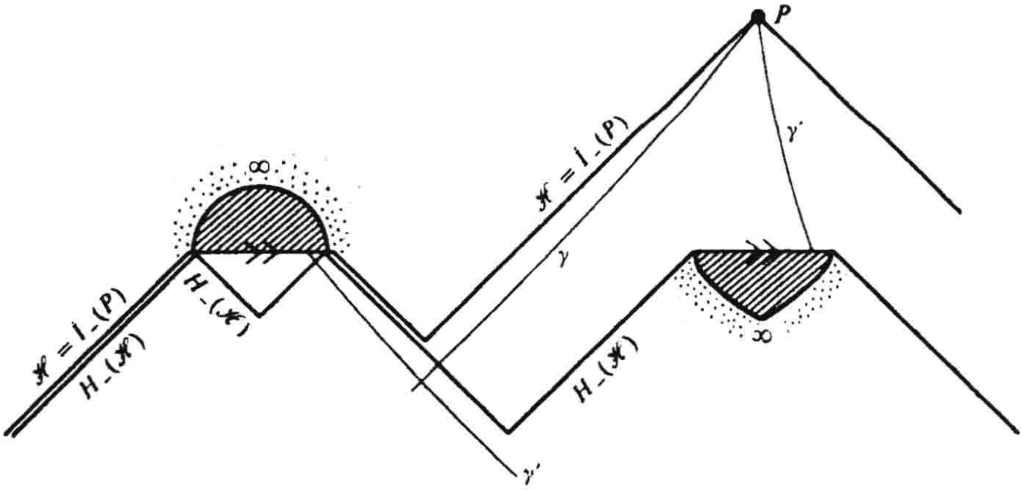


$\mathscr{H}$
γ
γ'
∞
$I_-[\mathscr{H}]$
$H_-(\mathscr{H})$
$I_-[\mathscr{H}]$

FIGURE 43. The type of situation to keep in mind for Theorem II.

in some compact region [*cf.* Fig. 33]. For Theorem III, an awkward type of possible situation is illustrated in Fig. 44 for which $H_-(\mathscr{H})$ can intersect $\mathscr{H}$. (It is the presence of such a possibility which requires the causality assumption (IIIa) to be made.)

Lemma V. If $\mathscr{H} = \bar{\mathscr{H}}$ is edgeless and semispacelike, then int $D_-(\mathscr{H})$ consists of all points $X \in I_-[\mathscr{H}]$ for which the set $J_+(X) \cap \bar{I}_-[\mathscr{H}]$ is both compact and contains no points at which strong causality fails.



P
$\mathscr{H} = \dot{I}_-(P)$
γ'
∞
γ
$\mathscr{H} = \dot{I}_-(P)$
$H_-(\mathscr{H})$
$H_-(\mathscr{H})$
$H_-(\mathscr{H})$
∞
γ'

identify along ⟶⟫⟶

FIGURE 44. An awkward possibility to bear in mind for Theorem III.

Write $\mathscr{X}$ for the set of such points X. [For $J_+(X)$, see (9.2).] Let $X \in \mathscr{X}$ and consider a future-inextendible timelike curve with X as past end-point. If this did not meet $\mathscr{H}$, it would, by the compactness of $\mathscr{Y} = J_+(X) \cap \bar{I}_-[\mathscr{H}]$, have to have an accumulation point in $\mathscr{Y}$. This is impossible since strong causality holds in $\mathscr{Y}$. Thus, $\mathscr{X} \subset D_-(\mathscr{H})$. Furthermore $X \notin H_-(\mathscr{H})$, since otherwise, by (10.16), a future-inextendible null geodesic with past end-point X would lie in $\mathscr{Y}$, $\mathscr{H}$ being edgeless. (Such a null geodesic could not be contained in a compact set unless strong causality failed in the set.) Hence $\mathscr{X} \subset \operatorname{int} D_-(\mathscr{H})$. For the converse, we recall, first, that if $Z \in \operatorname{int} D_-(\mathscr{H})$ then any future-inextendible *nonspacelike* curve through Z must meet $\mathscr{H}$. Now strong causality cannot fail at any point $Z \in \operatorname{int} D_-(\mathscr{H})$. (Otherwise, in the critical case, we could use the "Q" of (11.2) to get ZQ as a null geodesic along which strong causality fails and whose maximal extension would not meet $\mathscr{H}$.) Suppose, on the other hand, that $\mathscr{Z}_0 = J_+(Z_0) \cap \bar{I}_-[\mathscr{H}]$ fails to be compact for some $Z_0 \in \operatorname{int} D_-(\mathscr{H})$. Cover $\mathscr{Z}_0$ by a locally finite system of causally convex open sets $\mathscr{U}_i$ which are small enough to ensure that any point U in $\mathscr{U}_i$ is the center of a normal coordinate open ball $\mathscr{B}_U$ containing $\bar{\mathscr{U}}_i$. Let $Z_0 \in \mathscr{U}_{i_0}$. Since $\mathscr{Z}_0$ is noncompact it contains a sequence of points W_j with no accumulation point in $\mathscr{Z}_0$. A nonspacelike curve from Z_0 to W_j exists and meets $\dot{\mathscr{B}}_{Z_0}$ in V_j, say. Let V be an accumulation point of V_i. The nonspacelike geodesic segment Z_0V meets $\dot{\mathscr{U}}_{i_0}$ in a unique point $Z_1 \in \mathscr{U}_{i_1}$. Then $\mathscr{Z}_1 = J_+(Z_1) \cap \bar{I}_-[\mathscr{H}]$ contains a subsequence of the W_j without an accumulation point in $\mathscr{Z}_1$ and so also fails to be compact. Repeating the argument we obtain a sequence $Z_0 \prec Z_1 \prec Z_2 \prec \ldots$ which lie on a nonspacelike curve γ. From the locally finite nature of the $\mathscr{U}_i$ system it follows that the Z_i have no accumulation point, so γ is a future-inextendible nonspacelike curve through Z_0 not meeting $\mathscr{H}$. Thus $Z_0 \notin \operatorname{int} D_-(\mathscr{H})$, showing that $\operatorname{int} D_-(\mathscr{H}) \subset \mathscr{X}$. This completes the proof of Lemma V.

If $\mathscr{Q}$ and $\mathscr{R}$ are subsets of $\mathscr{M}$ or points of $\mathscr{M}$, define $d(\mathscr{Q}, \mathscr{R})$ to be the *least upper bound of the lengths of all timelike curves with past end-point* (*in*) $\mathscr{Q}$ *and future end-point* (*in*) $\mathscr{R}$. (Put $d(\mathscr{Q}, \mathscr{R}) = 0$ if there are no such curves.) In some cases we shall have $d(\mathscr{Q}, \mathscr{R}) = \infty$ (but for bounded subsets of Minkowski space-time, for example, it is always finite: a timelike straight line connecting two Minkowskian points has the *maximum* length among all timelike curves connecting the points).

Lemma VI. If $\mathscr{H} = \bar{\mathscr{H}}$ is edgeless and semispacelike, then $d(X, \mathscr{H})$ is finite, and attained for some geodesic γ from X to $\mathscr{H}$, whenever $X \in \operatorname{int} D_-(\mathscr{H})$.

By Lemma V, we know that $\mathscr{Y} = J_+(X) \cap \bar{I}_-(\mathscr{H})$ is compact, and can therefore be covered by a finite number of small open normal coordinate neighborhoods $\mathscr{B}_i$. Since strong causality holds in $\operatorname{int} D_-(\mathscr{H})$, we can arrange the boundaries of the $\mathscr{B}_i$'s to be causally convex. Write d_i for the least upper bound of lengths of timelike curves in $\mathscr{B}_i$. Then from local arguments it follows d_i is finite, and attained (by a geodesic, in fact). If γ is a timelike curve from X to $\mathscr{H}$ then the length of γ cannot exceed $d = \Sigma d_i$. Hence $d(X, \mathscr{H})$ is finite. Furthermore, from the compactness, it will follow that the maximum of $d(X, Y)$ for $Y \in \mathscr{H}$, will be *attained*, say, for $Y = Y_0$.

To construct γ of maximal length from X to $\mathscr{H}$ [so that γ actually is of length $d(X, Y)$], choose a normal coordinate ball $\mathscr{B}$, with center Y_0, and choose $Z = Z_0$ to maximize $d(X, Z) + d(Z, Y_0)$, where $Z \in \mathscr{Y}$ varies over the compact boundary $\dot{\mathscr{B}}$ of $\mathscr{B}$. Define γ to be the geodesic with future end-point Y_0 which contains Z_0 and extends into the past to a length d. Repeat the above argument, applying it to Z_0 in place of Y_0 and maximize $d(X, V) + d(V, Z_0)$. We get V_0, which must lie on γ (since a timelike curve with a corner can always be increased in length by smoothing the corner). Continuing the process, we see that γ must terminate with past end-point at X, establishing Lemma VI.

Lemma VII. If $\mathscr{H}$ is a spacelike hypersurface and γ is a timelike geodesic segment from X to $\mathscr{H}$ which maximizes $d(X, \mathscr{H})$, then γ meets $\mathscr{H}$ orthogonally and contains no point (other than possibly X) which is conjugate to $\mathscr{H}$.

The argument is not essentially different from that for a manifold with a positive definite metric (see Milnor [58]). A point conjugate to $\mathscr{H}$ on γ is a focal point for the congruence Γ of null geodesics which meet $\mathscr{H}$ orthogonally [that is, a point at which the θ of (7.40) becomes infinite].

We can also apply Lemma VII in a limiting case, to the hypersurface $\mathscr{H}$ of (11.6) required for Theorem III. In this case, if γ is a timelike geodesic segment from X to $\mathscr{H}$ which maximizes $d(X, \mathscr{H})$, then γ, to be "orthogonal" to $\mathscr{H}$, would have to contain P. Every other point of $\mathscr{H}$ lies on a null geodesic on $\mathscr{H}$ which extends into the future to points more distant from X. In this case, then, we interpret Lemma VII as saying that γ passes through P and contains no point (other than possibly X) which is conjugate to P (that is, which is a focal point of the congruence Γ of timelike or null geodesics through P).

We are now in a position to apply these results to Theorems II and III. Let t^a denote the unit (future-pointing) tangent vector field to the timelike γ's of Γ (in each case). By (IIb) or (IIIb), we have $\theta > 0$ at some point in $I_-[\mathscr{H}]$ on each $\gamma \in \Gamma$. Thus by (IIc) or (IIIc) and the Raychaudhuri effect [*cf.* (7.40)–(7.47); also (7.39)], we have a focal point $G(\theta = \infty;\ -\rho = \infty)$ somewhere on each $\gamma \in \Gamma$, to the past of $\mathscr{H}$. [This uses past-timelike completeness (IId), (IIId) and past-null completeness in (IIId).] If there is more than one such focal point, we choose the one nearest to $\mathscr{H}$ (and $\neq P$). Now

G will vary continuously with γ. By the compactness of $\mathscr{H}$ (Theorem II) or the fact (Theorem III) that the timelike and null directions at P constitute a compact system, we obtain the fact that the region $\mathscr{G}$ swept out by the sements of γ from G to $\mathscr{H}$ will be *compact*.

Now, by Lemmas VI and VII we have the result that $\operatorname{int} D_-(\mathscr{H}) \subset \mathscr{G}$, and thus $D_-(\mathscr{H}) \subset \mathscr{G} \cup \mathscr{H}$. In particular, $H_-(\mathscr{H}) \subset \mathscr{G} \cup \mathscr{H}$. In Theorem II, $\mathscr{H}$ is spacelike, so $H_-(\mathscr{H}) \cap \mathscr{H} = \emptyset$. The Cauchy horizon $H_-(\mathscr{H})$, being a closed subset of the compact set $\mathscr{G}$, is now *compact*. By (10.15) and (9.7), $H_-(\mathscr{H})$ is thus a compact C^0 manifold without boundary. By (IId), we can extend each γ into the past to length greater than $\max d(G, \mathscr{H})$. Then each γ meets $H_-(\mathscr{H})$ by Lemmas VI and VII. Put $F = \gamma \cap H_-(\mathscr{H})$ and define $p(F)$ to be the maximum of lengths of the segments of γ from F to $\mathscr{H}$. (Attained, for fixed F, because of the compactness of $\mathscr{H}$ and the Γ system.) It is not hard to see that $P(F)$ [with $F \in H_-(\mathscr{H})$] attains its *minimum* value. Let $F = F_0$ minimize $p(F)$. Now, a null geodesic η on $H_-(\mathscr{H})$ has F_0 as past end-point. Let F_1 be just to the future of F_0 on η. The length of the jointed curve from F_0 to $\mathscr{H}$, which consists of the small segment F_0F_1 of η together with a maximal γ curve from F_1 to $\mathscr{H}$, cannot be less than $p(F_0)$. Such γ exists since $D_-(\mathscr{H}) \subset \mathscr{G}$. Smoothing the corner, and moving the curve just away from $H_-(\mathscr{H})$, we get a timelike curve ζ from F_0 to $\mathscr{H}$ of length k greater than $p(F_0)$. Taking B on ζ very close to F_0, so $d(B, \mathscr{H}) > p(F_0)$ and letting B approach F_0, we get a limiting position of the (relevant) γ through B (from the compactness of $\mathscr{H}$). This would give us $p(F_0) \geqslant k > p(F_0)$ which is the required contradiction establishing Theorem II.

In Theorem III, $\mathscr{H}$ is not spacelike, so $H_-(\mathscr{H})$ might intersect $\mathscr{H}$ (Fig. 44). Thus, we cannot assume that $H_-(\mathscr{H})$ is compact and the above argument will break down. Instead, we bring in the strong causality. Define a relation $\langle$ on $H_-(\mathscr{H})$ by

$$U \langle V \text{ IF AND ONLY IF } I_-(U) \subset I_-(V) \tag{11.7}$$

[that is, $U \in \bar{I}_-(V)$]. Then $\langle$ is an order relation on $H_-(\mathscr{H})$ (weaker than $\prec$). Put $\mathscr{A} = \mathscr{G} \cap H_-(\mathscr{H})$. Then $\mathscr{A}$ is compact $H_-(\mathscr{H})$ being closed and $\mathscr{G}$ compact. Set $\mathscr{K} = \mathscr{A} \cap \dot{\mathscr{G}} \subset H_-(\mathscr{H}) - \operatorname{int}\mathscr{G} \subset \mathscr{H}$. Let $U \subset \mathscr{A}$. Then a future-inextendible null geodesic α on $H_-(\mathscr{H})$ has past end-point U. By compactness and strong causality, α must leave $\mathscr{A}$ at $U_0 \in \mathscr{K}$, say, and we shall have $U \langle U_0$. But in the neighborhood of $\alpha - \mathscr{A}$ are timelike curves leading to P [since $\alpha - \mathscr{A} \subset \mathscr{H} = \dot{I}_-(P)$]. These timelike curves must meet $H_-(\mathscr{H}) - \mathscr{H} \subset \mathscr{A}$ on their way to P. In the limit as such curves approach α, this intersection point has an accumulation point V on $\mathscr{A}$ and we readily verify $U_0 \langle V$ (with $U_0 \neq V$ because of strong causality). Similarly V gives rise to V_0 on $\mathscr{K}$ with $V \langle V_0$, and to W with $V_0 \langle W (V_0 \neq W)$, etc. The sequence of distinct points $U_0 \langle V_0 \langle W_0 \langle \ldots$ on $\mathscr{K}$ must have an accumulation point, by the compactness of $\mathscr{K}$. We would have to have a breakdown of strong causality at this point. This establishes Theorem III.

One of the significant features of Theorem III is that it is not necessary to make assumptions apart from strong causality as to the nature of the universe as a whole. The theorem refers only to the past null cone of P and its interior, these being regions in principle observable by an observer at P. In time-reversed form the theorem applies to sufficiently uniform collapse situations (although not quite of the generality of those treated in Theorem I) and requires no information of a global cosmological nature. (We choose P at the center and assume the collapse is similar to a late stage of the Oppenheimer–Snyder situation.) Also, since no Cauchy hypersurface condition is needed, one is led to believe that even if the Cauchy horizon of Fig. 41 still exists in generic situations, this would probably not enable us to avoid singularities elsewhere.

Of course, when applied to the past cone of P, we would not expect the divergence of the geodesics to change sign until a substantial portion of the entire universe has been encompassed. To estimate where this point should occur, one would need to know the density of material in the universe and knowledge of this is very uncertain at present. Also, much of the matter, for example, galaxies, is very irregularly distributed, so the total effect might be somewhat complicated. It ceases to be clear whether (IIIb) would be satisfied at all, by virtue merely of observed material objects. The idea of Hawking and Ellis [44], however, is to use the *radiation* density of the universe, whose present existence is in the form of a highly isotropic distribution with a spectrum apparently that of a blackbody at 3^0 absolute. They argue that this radiation density, owing to its uniformity, ought in *itself* to be sufficient ultimately to ensure that (IIIb) is satisfied, except in the unlikely possibility of there being a large amount of ionized gas present. If this gas is present, then *it* should be sufficient to produce the desired focusing instead.

The indications are, then, that if Einstein's equations hold, then (curvature?) singularities[34] are a real feature of our universe. (Actually, Einstein's equations are not very strongly used. For example, if we were to replace Einstein's theory by the Brans–Dicke [13] theory, then the conclusions of Theorems I, II, and III would be essentially unaffected.) More results are needed, however, if we are to get any feeling for the structure of these peculiar (presumably highly curved) regions of space-time. And really to ascertain the kind of physics that might be involved in such a region, we would need a much deeper understanding of the interrelation between matter, quantum theory, and space-time structure than we have at present.

---

[34] By a "singularity" I mean a peculiar region, in which the local physics would be drastically affected, possibly to the extent that the smooth manifold picture of space-time would no longer be adequate (*cf.* discussion in Section 1).

Under normal circumstances, general relativity can, for practical purposes, remain remarkably apart—almost aloof—from the rest of physics. At a space-time singularity, the very reverse must surely be the case!

REFERENCES

1. Y. Aharonov and L. Susskind, *Phys. Rev.* **158**, 1237 (1967).

2. R. A. Alpher and R. C. Herman, *Nature* **162**, 774 (1948).

3. A. Avez, *Compt. Rend.* **254**, 3984 (1962).

4. A. Avez, *Inst. Fourier* **105** (1963).

5. J. Bardeen, to be published.

6. C. Behr, *Zeit fur Astrophysik* **60**, 286 (1965).

6a. P. G. Bergmann, *Phys. Rev. Lett.* **12**, 139 (1964).

7. B. Bertotti, *Proc. Roy. Soc.* (*Lond.*) **A294**, 195 (1966).

8. H. Bondi, *Nature*, **186**, 535 (1960).

9. H. Bondi, in *Lectures on General Relativity*: *1964 Brandeis Summer Institute in Theoretical Physics, Vol. I* (Prentice-Hall, Englewood Cliffs, 1965).

10. H. Bondi, F. A. E. Pirani, and I. Robinson, *Proc. Roy. Soc.* (*Lond.*) **A251**, 519 (1959).

11. H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, *Proc. Roy. Soc.* (*Lond.*) **A269**, 21 (1962).

12. R. H. Boyer and R. W. Lindquist, *J. Math. Phys.* **8**, 265 (1967).

13. C. Brans and R. H. Dicke, *Phys. Rev.* **124**, 925 (1961).

14. D. Brill, *Annals of Phys.* **7**, 466 (1959).

14a. D. Brill and S. Deser, *Phys. Rev. Lett.* **20**, 75 (1968).

15. H. W. Brinkmann, *Nat. Acad. Sci.* (U.S.) **9**, 1 (1923).

16. H. A. Buchdahl, *Nuovo Cim.* **10**, 96 (1958); **25**, 486 (1962).

17. E. Calabi and L. Markus, *Ann. Math.* **75**, 63 (1962).

18. B. Carter, *Phys. Rev.* **141**, 1242 (1966).

19. B. Carter, *Stationary Axi-symmetric Systems in General Relativity* (Ph.D. Dissertation, Cambridge University, 1967).

20. E. Cartan, *Ann. Ec. Norm. Sup.* **40**, 325 (1923); **41**, 1 (1924).

21. S. Chandrasekhar, *M. N.* **95**, 207 (1935).

21a. S. Chandrasekhar, *Proc. Nat. Acad. Sci.* **57**, (1967).

22. G. F. Chew, *Sci. Prog.* **51**, 529 (1963).

23. R. Debever, *Comptes Rend.* **249**, 1324, 1744 (1959).

23a. V. de la Cruz and W. Israel, *Nuovo Cim.* **51A**, 745, (1967).

24. R. H. Dicke, *Phys. Rev.* **125**, 2163 (1962).

25. R. H. Dicke, P. J. E. Peebles, P. G. Roll, and D. T. Wilkinson, *Ap. J.* **142**, 414 (1965).

26. A. G. Doroshkevich, Ya. B. Zel'dovich and I. D. Novikov, *Zhur. Eksp. Teor. Fiz.* **49**, 170 (1965); Engl. tr.: *Sov. Phys.—JETP.* **22**, 122 (1966).

27. A. S. Eddington, *Nature* **113**, 192 (1924).

28. J. Ehlers and W. Kundt, in *Gravitation* (ed. L. Witten; John Wiley and Sons, New York 1962).

29. A. Einstein, *Ann. Physik.* **94**, 769 (1916).

30. A. Einstein and J. Grommer, *S. B. Preuss. Akad. Wiss.* **2**, (1927).

31. A. Einstein, L. Infeld, and B. Hoffmann, *Ann. Math.* **39**, 65 (1938).

32. D. Finkelstein, *Phys. Rev.* **110**, 965 (1958).

33. G. Gamow, *Nature* **162**, 680 (1948).

34. R. P. Geroch, *J. Math. Phys* **8**, (1967).

35. R. P. Geroch, *Singularities in the Spacetime of General Relativity* (Ph.D. Dissertation, Princeton University, 1967).

36. J. N. Goldberg, in *Gravitation* (ed. L. Witten; John Wiley and Sons, New York, 1962).

37. J. C. Graves and D. R. Brill, *Phys. Rev.* **120**, 1507 (1960).

38. Haefliger; see *Topics on Space-Times* by A. Lichnerowicz, Ch. V, this volume.

39. B. K. Harrison, K. S. Thorne, M. Wakano, and J. A. Wheeler, *Gravitation Theory and Gravitational Collapse* (University of Chicago Press, Chicago, 1965).

40. S. W. Hawking, *Proc. Roy. Soc. (Lond.)* **A294**, 511 (1966).

41. S. W. Hawking, *Proc. Roy. Soc. (Lond.)* **A295**, 490 (1966).

42. S. W. Hawking, *Proc. Roy. Soc. (Lond.)* **A300**, 187 (1967).

43. S. W. Hawking and G. F. R. Ellis, *Physics Letters*, **17**, 246 (1965).

44. S. W. Hawking and G. F. R. Ellis, *Ap. J.* in press (1968).

45. S. Helgason, *Differential Geometry and Symmetric Spaces* (Academic Press, New York, 1962).

46. L. Infeld and B. L. van der Waerden, *S. B. Preuss Akad.* **9**, 380 (1933).

47. P. Jordan, J. Ehlers, and R. K. Sachs, *Akad. Wiss., Mainz* **1** (1961).

47a. R. Kantowski, and R. K. Sachs, *J. Math. Phys.* **7**, 443 (1966).

48. R. P. Kerr, *Phys. Rev. Lett.* **11**, 237 (1963).

49. A. Komar, *Phys. Rev.* **104**, 544 (1956).

50. E. Kronheimer and R. Penrose, *Proc. Camb. Phil. Soc.* **63**, 481 (1967).

50a. M. D. Kruskal, *Phys. Rev.* **119**, 1943 (1960).

51. N. H. Kuiper, *Ann. of Math.* **50**, 916 (1949).

52. W. Kundt, *Z. Phys.* **172**, 488 (1963).

53. E. M. Lifshitz and I. M. Khalatnikov, *Advan. in Phys.* **12**, 185 (1963).

54. R. W. Lindquist and J. A. Wheeler, *Revs. Mod. Phys.* **29**, 432 (1957).

54a. S. MacLane and G. Birkkoff, *Algebra* (Macmillan, New York, 1967).

55. L. Markus, *Ann. Math.* **62**, 411 (1955).

56. R. F. Marzke and J. A. Wheeler, in *Gravitation and Relativity* (ed. H. Y. Chiu and W. F. Hoffmann; Benjamin, New York, 1964).

57. J. W. Milnor, *Enseignement Math.* (2) **9**, 198 (1963).

58. J. Milnor, *Morse Theory* (Annals of Math. Studies, Princeton University Press, Princeton, 1963).

59. C. W. Misner, *J. Math. Phys.* **4**, 924 (1963).

60. C. W. Misner, *Ap. J.* in press (1968).

61. C. W. Misner and A. H. Taub, to be published.

62. C. Møller, *The Theory of Relativity* (Oxford Univ. Press, Oxford, 1952).

63. E. T. Newman, L. Tamburino, and T. Unti, *J. Math Phys.* **4**, 915 (1963).

64. E. T. Newman and R. Penrose, *J. Math. Phys.* **3**, 566 (1962); **4**, 998 (1963).

65. K. Nomizu, *Lie Groups and Differential Geometry* (Herald Printing Co. Ltd., Tokyo, 1956).

66. G. Nordström, *Ann. der Physik* **42**, 533 (1913).

66a. I. V. Novikov, *J E T P. Lett.* 3, 142 (1966).

67. J. R. Oppenheimer and G. Volkoff, *Phys. Rev.* **55**, 374 (1939).

68. J. R. Oppenheimer and H. Snyder, *Phys. Rev.* **56**, 455 (1939).

69. J. P. Ostriker, P. Bodenheimer and D. Lynden-Bell, *Phys. Rev. Lett.* **17**, 816 (1966).

70. W. T. Payne, *Am. J. Phys.* **20**, 253 (1952).

71. R. Penrose, *Proc. Camb. Phil. Soc.* **55**, 137 (1959).

72. R. Penrose, *Ann. of Phys.* **10**, 171 (1960).

73. R. Penrose, in P. G. Bergmann's Aeronautical Research Lab. Tech. Documentary Rept. 63-56: *Quantization of Generally Covariant Fields* (Office of Aerospace Research, U.S. Air Force, 1963).

74. R. Penrose, in *Relativity, Groups, and Topology* (ed. C. M. DeWitt and B. S. DeWitt; Gordon and Breach, New York, 1964).

75. R. Penrose, *Phys. Rev. Letters* **14**, 57 (1965).

76. R. Penrose, *Proc. Roy. Soc.* (*Lond.*) **A284**, 159 (1965).

77. R. Penrose, *Rev. Mod. Phys.* **37**, 215 (1965).

78. R. Penrose, in *Perspectives in Geometry and Relativity* (Indiana University Press, Bloomington, 1966).

79. R. Penrose, *J. Math. Phys.* **8**, 345 (1967).

80. R. Penrose, *Intern. J. Theor. Phys.* **1**, (1968).

80a. A. A. Penzias and R. W. Wilson, *Ap. J.* **142**, 419 (1965).

80b. A. Z. Petrov, *Sci. Not., Kazan State University* **114**, 55 (1954).

81. F. A. E. Pirani, *Phys. Rev.*, **105**, 1089 (1957).

82. F. A. E. Pirani, in *Lectures on General Relativity: 1964 Brandeis Summer Institute in Theoretical Physics Vol. I* (Prentice-Hall, Englewood Cliffs, 1965).

83. F. A. E. Pirani and A. Schild, in *Perspectives in Geometry and Relativity* (Indiana University Press, Bloomington, 1966).

84. J. Plebanski, *Acta Phys. Polon.* **27**, 361 (1965).

85. R. V. Pound and G. A. Rebka, *Phys. Rev. Letters* **4**, 337 (1960).

86. A. K. Raychaudhuri, *Phys. Rev.* **98**, 1123 (1955).

87. S. Refsdal, *Mon. Not. R. Astr. Soc.* **128**, 295 (1964); **128**, 307 (1964).

88. W. Rindler, *Mon. Not. Roy. Astr. Soc.* **116**, 6 (1956).

89. I. Robinson, *J. Math. Phys.* **2**, 290 (1961).

90. I. Robinson, A. Schild, and E. L. Schucking, eds.: *Quasi-Stellar Sources and Gravitational Collapse* (University of Chicago Press, Chicago, 1965).

91. H. S. Ruse, *Proc. Lond. Math. Soc.* **50**, 75 (1948).

92. H. Rudberg, *The Compactification of a Lorentz Space* (Thesis, University of Uppsala, Uppsala, Sweden, 1958).

93. R. K. Sachs, *Proc. Roy. Soc.* (*Lond.*) **A264**, 309 (1961).

94. R. K. Sachs, in *Recent Developments in General Relativity* (PWN/Pergamon, Warsaw/New York, 1962).

95. R. K. Sachs, *Proc. Roy. Soc.* (*Lond*) **A270**, 103 (1962).

96. A. Schild, *Am. J. Phys.* **28**, 778 (1960).

96a. J. A. Schouten, *Ricci-Calculus* (Springer, Berlin, 1954).

97. E. Schrodinger, *Expanding Universes* (Cambridge Univ. Press, Cambridge, 1956).

98. H. J. Seifert, *Naturforsch.* **22a**, 1356 (1967).

99. L. C. Shepley, *Proc. Nat. Acad. Sci.* **52**, 1403 (1964).

100. B. Spain, *Tensor Calculus* (Oliver and Boyd, London, 1960).

101. N. Steenrod, *The Topology of Fibre Bundles* (Princeton Univ. Press, Princeton, 1951).

102. K. Stellmacher, *Math. Annalen* **123**, 34 (1951).

103. J. L. Synge, *Relativity: the General Theory* (North-Holland, Amsterdam, 1960).

104. L. A. Tamburino and J. Winicour, *Phys. Rev.* **150**, 1039 (1966).

105. A. H. Taub, *Ann. Math.* **53**, 472 (1951).

106. J. Terrel, *Phys. Rev.* **116**, 1041 (1959).

107. K. S. Thorne, Relativistic Stellar Structure and Dynamics in 1966, *Les Houches Lectures* (ed. C. M. DeWitt, Gordon and Breach, New York, 1967).

108. A. Trautman, in *Lectures on General Relativity: 1964 Brandeis Summer Institute in Theoretical Physics, Vol. I* (Prentice-Hall, Englewood Cliffs, 1965).

109. A. Trautman, in *Perspectives in Geometry and Relativity* (IndianaUniversity Press, Bloomington, 1966).

110. B. L. van der Waerden, *Nachr. Ges. Wiss. Göttingen* **100** (1929).

111. H. Weyl, *Nachr Göttingen.* **99** (1921).

112. J. A. Wheeler, *Geometrodynamics* (Academic Press, New York, 1962).

113. J. A. Wheeler, in *Relativity, Groups and Topology* (ed. C. M. DeWitt and B. S. DeWitt; Gordon and Breach, New York, 1964).

113a. H. Whitney, *Bull. Amer. Math. Soc.* **43**, 785 (1937).

114. E. T. Whittaker, *Proc. Roy. Soc.* (*Lond.*) **A158**, 38 (1937).

115. L. Witten, *Phys. Rev.* **113**, 357 (1959).

37

TWISTOR QUANTIZATION AND CURVED SPACE-TIME

International Journal of Theorectical Physics **1(1)** (1968), 61–99

This article launched various new developments in twistor theory, moving the subject significantly more into physics than would have been apparent from those bits of algebra and geometry described in Chapter 29. We have the first appearance of a contour-integral formula for the twistor representation of solutions of the massless free-field equations (although the delayed article of Chapter 39 was actually written roughly two years in advance of the present one). There is the first indication, here, of how gravitation might distort twistor space (compare Chapters 67, 70, 71), where a Hamiltonian appears involving a holomorphic twistor function of homogeneity degree 2. We have the twistor-space simplectic structure and the first indications of twistor quantization rules; also the scalar product beween twistor functions appears, with some of its attendant difficulties.

Twistor Quantisation and Curved Space-Time†

ROGER PENROSE

Department of Mathematics, Birkbeck College, University of London, London, England

Abstract

The formalism of twistors [the 'spinors' for the group O(2, 4)] is employed to give a concise expression for the solution of the zero rest-mass field equations, for each spin ($s = 0, \frac{1}{2}, 1, \ldots$), in terms of an arbitrary complex analytic function $f(Z^\alpha)$ (homogeneous of degree $-2s - 2$). The four complex variables Z^α are the components of a twistor. In terms of twistor space (C-picture) it is analytic structure which takes the place of field equations in ordinary Minkowski space-time (M-picture). By requiring that the singularities of $f(Z^\alpha)$ form a disconnected pair of regions in the upper half of twistor space, fields of positive frequency are generated.

The twistor formalism is adapted so as to be applicable in curved space-times. The effect of conformal curvature in the M-picture is studied by consideration of plane (-fronted) gravitational 'sandwich' waves. The C-picture still exists, but its complex structure 'shifts' as it is 'viewed' from different regions of the space-time. A weaker symplectic structure remains. The shifting of complex structure is naturally described in terms of Hamiltonian equations and Poisson brackets, in the twistor variables Z^α, $\bar{Z}_\alpha$. This suggests the correspondence $\bar{Z}_\alpha = \partial/\partial Z^\alpha$ as a basis for quantization. The correspondence is then shown to be, in fact, valid for the Hilbert space of functions $f(Z^\alpha)$, which give the above twistor description of zero rest-mass fields. For this purpose, the Hilbert space scalar product is described in (conformally invariant) twistor terms. The twistor expressions for the charge and the mass, momentum and angular momentum (both in 'inertial' and 'active' versions, in linearised theory) are also given.

It is suggested that twistors may supply a link between quantum theory and space-time curvature. On this view, curvature arises whenever a 'shift' occurs in the interpretation of the twistor variables Z^α, $\bar{Z}_\alpha$ as the twistor 'position' and 'momentum' operators, respectively.

1. *Introduction*

In an earlier paper (Penrose, 1967a) the formalism of *twistor algebra* was developed, which treats the geometry of Minkowski space-time from the point of view of its null line and null cone structure. In the present paper this formalism is used to give a concise description of zero rest-mass fields, and the formalism is developed further so as to be applicable in curved space-times as well as flat. An unexpected interconnection between these two ideas leads to a new view of the relation between quantum theory and space-time curvature.

---

† This work was partly carried out during the author's five-month stay at Cornell University.

According to the twistor formalism, any conformally covariant operation in Minkowski space-time has a description in purely twistor terms, and with the introduction of a fixed skew-symmetric 'metric' twistor $I_{\alpha\beta}$ we can also express Poincaré covariance in purely twistor terms. The twistor algebra leads to a geometrical picture of phenomena (the C-picture) which, although strikingly different from the usual space-time description (M-picture), is nevertheless completely equivalent to it. The points (i.e. 'events') of the M-picture correspond to non-local structures ('lines'—each with the topology S^2) in the C-picture; conversely the points of the C-picture correspond to non-local structures in the M-picture (to null straight lines or, more generally, to certain twisting null line systems). An outline of the results we require here will be given in Section 2.

The motivation for rewriting physical quantities in twistor terms springs from several directions. In the first instance, there is simply the hope that when a formalism so different from the usual one is used, new insights may be gained. While it is true that certain important concepts, which were easy to express in the old formalism, can become complicated in the new (which need not be a serious drawback, since the old formalism is always at hand when required), there are other operations of great utility in the new formalism which one would be unlikely to come upon solely by considerations with the old formalism. But are these new operations likely to be of any particular importance to physics? It is here that I must be more specific and mention some of those features which motivate the specific choice of a twistor formalism for the description of space-time.

One of these features is that the twistor formalism is the natural vehicle for the *algebraic* description of *conformal invariance*. Twistors are, in fact, the 'spinors' appropriate to the six-dimensional pseudo-orthogonal group O(2,4), which is 2-1 isomorphic with the 15-parameter conformal group of (the compactified) Minkowski space-time (Cartan, 1914; Brauer & Weyl, 1935; Hepner, 1962; Murai, 1953, 1954, 1958; Segal, 1967). The connected component of the twistor group SU(2,2) is 2-1 isomorphic with connected component of O(2,4) so that twistors $\left(\text{of valence } \begin{bmatrix}1\\0\end{bmatrix}\right)$ give a four-valued four-dimensional irreducible representation of the restricted conformal group. Any other finite-dimensional representation of the conformal group can be expressed as a direct sum of twistor representations $\left(\text{of general valence } \begin{bmatrix} p \\ q \end{bmatrix}\right)$.

But is there reason to believe that the conformal group has any fundamental significance to physics?† The attitude adopted here will be that fields of zero rest-mass (which are conformally invariant) have primary significance; and that in some way, rest-mass emerges as a feature of interactions between these primary zero rest-mass fields. A simple way such an interaction might be expressed emerges in the van der Waerden description (van der Waerden, 1929) of the Dirac equation:

$$\nabla^{B'A}\phi_A = \mu\psi^{B'}, \qquad \nabla_{B'A}\psi^{B'} = -\mu\phi_A \tag{1.1}$$

Here μ is a real constant, $2^{1/2}\mu\hbar$ being the mass of the field. We may regard (1.1) as describing two neutrino-like fields ϕ_A, $\psi^{B'}$ whose free-field equations are given by putting $\mu = 0$ in (1.1). These free-field equations are then conformally invariant. We may take the view that μ is simply a coupling constant, given for all time, and that the two-field interaction (1.1) simply breaks this conformal invariance. The conformal group is then only strictly a symmetry of very high energy physics—where energies are high enough that the rest-mass interaction may be neglected (Kastrupp, 1962, 1966). Alternatively we may imagine that μ is 'really' a new variable field which for some reason (perhaps of a cosmological nature or from stability considerations, say) 'happens' to be transformable to a constant with a very high degree of accuracy. With further equations on μ, the entire set of equations can be made conformally invariant, the equations (1.1) referring to a *three*-field conformally invariant interaction. Finally, we may take the view that the interaction terms in (1.1) are of a phenomenological nature and the 'accurate' equations are really much more involved. This view is presumably what would be implied by the renormalization procedure. In addition, it is conceivable that general relativity has some significant role to play in connection with the existence of rest-mass, since gravitation is the only (known) phenomenon of nature which requires that a definite choice be made for the zero of energy, namely that it should coincide with the zero of active mass.

I do not wish to prejudice the issue here as to the ultimate nature of rest-mass. The attitude is only that it should be of significance to talk about free fields and that such fields may be viewed as having zero rest-mass. A description in terms of twistors should then have some importance as a 'background' formalism. Rest-mass could then be treated along with other interactions at a later stage.

---

† For a discussion of the relation of the conformal group to physics, see Fulton, T., Rohrlich, F. and Witten, L. (1962). *Rev. Mod. Phys.*, **34**, 442.

A second feature of twistor analysis, which has been highly instrumental in the motivation for its original development, lies in the extent to which it 'geometrises' an important aspect of quantum mechanics, namely the splitting of field amplitudes into positive and negative frequency parts. In the twistor formalism, instead of resorting to Fourier analysis, it is possible to exploit an alternative description based on the positioning of singularities of analytically extended functions.† As we shall see shortly, a positive or negative frequency field will arise according as the singularities of a certain analytic (holomorphic) function representing the field, form a disconnected pair of regions in the upper or the lower half of twistor space. It is the fact that the twistors lead to a 'mild' form of 'complexification' of the space-time which enables this idea to take on a 'geometrical' significance. The geometry of the C-picture involves its *complex analytic* structure. The analytic nature of functions defined in twistor space then yields the entire structure of zero rest-mass fields in the M-picture; in particular, field equations and time-development, now become simply particular aspects of C-picture analyticity.

Some of these matters will be described in Section 3. The aim there is to show that zero rest-mass fields find a very natural and remarkably *simple* description in twistor terms. This reinforces a belief that twistors might possibly occupy a position in physics of deeper importance than just as a technical device. But if twistors really do occupy such a position, it would have to be possible to overcome one of the supreme obstacles to such a viewpoint, namely that the formalism would in principle, at least, have to be applicable in (conformally) *curved* space-times. For even if general relativity is not the correct theory of gravitation (and it is, to say the very least, the best theory of gravitation available at the present time), there can be little doubt that the conformal structure of space-time, as defined by its null cones, *does differ* from that of Minkowski space-time. (A critical feature of theories of gravitation based on a conformally flat null-cone structure, e.g. Nordström's theory, is that there is *no* resultant bending of light by a massive body. It may be taken that this, at least, is experimentally disproved.)

The essential difficulty involved in attempting to adapt the twistor formalism to conformally curved space-time lies in the fact that the *complex analytic structure* of the C-picture is *destroyed by conformal curvature* of the M-picture (Section 4). This would seem, at first sight, to invalidate completely the use of twistors for conformally curved manifolds, since it is precisely the complex analytic structure of the C-picture which gives rise to all its important properties. However, it is here that twistors supply an unexpected *link between quantum mechanics and space-time curvature*. A non-analytic transformation of twistor space is one which mixes up the twistor coordinates Z^α with their complex conjugates $\bar{Z}_\alpha$. But the precise type of non-analytic transformation of the C-picture which is induced by the presence of conformal curvature in the M-picture turns out to be one preserving Poisson brackets, where $\bar{Z}_\alpha$ is regarded as the canonical conjugate of Z^α. This suggests that in the passage to quantum theory, Z^α should be regarded as an operator, where the operator $\bar{Z}_\alpha$ is identified with $\partial/\partial Z^\alpha$. (Planck's constant will be absorbed into the definition of the twistor variables Z^α.) With this identification, one can still operate with analytic functions in twistor space. What in the 'classical' theory was a 'canonical' non-analytic transformation of twistor space, corresponds in the quantized (one-particle) theory, to linear transformation of the *space of analytic functions* defined on twistor space. In fact, with the appropriate choice of norm, these become unitary transformations of a Hilbert space. The operations Z^α and $\partial/\partial Z^\alpha$ now become explicit operations in the M-picture which apply to zero rest-mass fields. The operator Z^α lowers the spin by $\frac{1}{2}\hbar$, while $\partial/\partial Z^\alpha$ raises it by $\frac{1}{2}\hbar$. The identification of $\bar{Z}_\alpha$ with $\partial/\partial Z^\alpha$ becomes explicitly consistent with the Hilbert space scalar product. Thus, we must regard the analytic structure of the C-picture as being relevant to the *quantum* structure of fields (e.g. to one-particle states), while the non-analytic transformations induced by curvature in the M-picture refer essentially to the *classical limit*.

---

† See, for example, Streater, R. F. and Wightman, A. S. (1964). *PCT, Spin and Statistics and All That*. W. A. Benjamin, New York.

We are led to a new view of the nature of space-time curvature. Let us adopt the attitude that in a certain region of space-time, the twistor variables Z^α are the ones, in terms of which physical quantities are to be expressed. The Z^α are the C-space 'position' operators and the $\bar{Z}_\alpha = \partial/\partial Z^\alpha$ are the conjugate 'momenta'. As we pass through a region of conformal curvature (e.g. a gravitational wave) the 'position' and 'momentum' variables get mixed up. We may take a 'self-consistent' or 'Machian' view of the correspondence principle, whereby the particular choice of quantum variables which are to be regarded as 'position' variables is governed by the nature of the large-scale physical structures in the region under consideration. Then, as we move to another region of space-time, the influence of a change in the large-scale structures is to effect a 'shift' in the natural interpretation of the variables. Thus, curvature of space-time emerges as a phenomenon intimately connected with such an attitude to the correspondence principle and with the quantum structure of nature.

2. *Outline of Twistor Algebra*

The main features of twistor algebra that will be required here will now be reviewed. For further details, the reader is referred to an earlier paper (Penrose, 1967a). Let x^0, x^1, x^2, x^3 be standard Minkowski coordinates in flat space-time M—metric given by

$$ds^2 = (dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2$$

Introduce a 2-spinor notation,† relating $x^{AA'}$, $x_{AA'}$ to x^a by

$$x^{00'} = x_{11'} = 2^{-1/2}(x^0 + x^1), \qquad x^{01'} = -x_{10'} = 2^{-1/2}(x^2 + ix^3) \qquad (2.1)$$

$$x^{10'} = -x_{01'} = 2^{-1/2}(x^2 - ix^3), \qquad x^{11'} = x_{00'} = 2^{-1/2}(x^0 - x^1)$$

When the coordinates x^a are real, the matrices $(x^{AA'})$ and $(x_{AA'})$ are Hermitian so we have two real coordinates u, v and one complex coordinate ζ, given by

$$u = x^{00'}, \qquad v = x^{11'}, \qquad \zeta = x^{01'}, \qquad \bar{\zeta} = x^{10'} \qquad (2.2)$$

The metric of M now takes the form

$$ds^2 = 2\,du\,dv - 2\,d\zeta\,d\bar{\zeta} \qquad (2.3)$$

A twistor Z^α, of valence $\begin{bmatrix}1\\0\end{bmatrix}$ is a quantity with four complex components Z^0, Z^1, Z^2, Z^3, its complex conjugate $\bar{Z}_\alpha$ being a twistor of valence $\begin{bmatrix}0\\1\end{bmatrix}$, related to Z^α by

$$\bar{Z}_0 = \overline{Z^2}, \qquad \bar{Z}_1 = \overline{Z^3}, \qquad \bar{Z}_2 = \overline{Z^0}, \qquad \bar{Z}_3 = \overline{Z^1} \qquad (2.4)$$

(When the bar extends only over the kernel symbol, this refers to the twistor complex conjugation operation. When the bar extends over both the kernel symbol and its index, then this means simply the complex conjugate of the complex number involved.) The twistor Z^α

---

† The primed index letter A' must be regarded as a distinct letter from A, so that no contraction is implied in $X^{AA'}$. Under complex conjugation, unprimed indices become primed and primed indices become unprimed.

is called *right-handed*, *left-handed*, or *null* respectively, according as the scalar

$$Z^{\alpha}\bar{Z}_{\alpha} = Z^{0}\overline{Z^{2}} + Z^{1}\overline{Z^{3}} + Z^{2}\overline{Z^{0}} + Z^{3}\overline{Z^{1}} \tag{2.5}$$

is positive, negative, or zero. (The summation convention will be employed throughout.) A null twistor Z^{α} describes a *null straight line* Z in M according to the scheme:

$$Z^{0}:Z^{1}:Z^{2}:Z^{3} = du:d\bar{\zeta}:i\zeta\,d\bar{\zeta} - iv\,du:i\bar{\zeta}\,du - iu\,d\bar{\zeta} \tag{2.6}$$

where u, v, ζ are coordinates [cf. (2.2), (2.1)] of some point P on Z and $du:dv:d\zeta$ define the direction of Z. Since Z is a null line we have, by (2.3),

$$du\,dv = d\zeta\,d\bar{\zeta} \tag{2.7}$$

Hence Z^{α}, being invariant under $x^{a} \to x^{a} + k\,dx^{a}$, is independent of the choice of P on Z. By (2.7), we can also write

$$Z^{0}:Z^{1}:Z^{2}:Z^{3} = d\zeta:dv:i\zeta\,dv - iv\,d\zeta:i\bar{\zeta}\,d\zeta - iu\,dv \tag{2.8}$$

Note that it is Z^{α} *up to a complex factor of proportionality* which uniquely corresponds to the line Z. When $Z^{0} = Z^{1} = 0$, we do not get a finite line in M, but rather a generator of the *null cone at infinity* for M. If we admit these lines *at infinity* as part of the conformal structure of M ('compactified' Minkowski space-time), then a null twistor Z^{α} (up to proportionality) only fails to define a unique null line in M if $Z^{\alpha} = 0$.

If X^{α} and Y^{α} are null twistors, defining null lines X and Y in M, respectively, then the condition for X and Y to meet (possibly at infinity) is

$$X^{\alpha}\bar{Y}_{\alpha} = 0 \tag{2.9}$$

(This is, of course, the same as $Y^{\alpha}\bar{X}_{\alpha} = 0$.) If X and Y do meet, at P, say, then any twistor Z^{α} which represents a generator of the *null cone* with vertex P has the form

$$Z^{\alpha} = \lambda X^{\alpha} + \mu Y^{\alpha} \tag{2.10}$$

This is necessarily null, by (2.9). If X and Y 'meet at infinity' (but do not both lie at infinity) then the null cone becomes a null hyperplane, with vertex P on the null cone at infinity. If both X and Y lie at infinity, then P becomes the vertex, I, of the null cone at infinity itself. Thus any point P in M, including those which lie at infinity, can be represented in twistor terms by a linear set of the type of (2.10). In fact we can use the twistor $P^{\alpha\beta} = X^{\alpha}Y^{\beta} - Y^{\alpha}X^{\beta}$, of valence $\begin{bmatrix}2\\0\end{bmatrix}$, to represent such a linear set, if desired. We can even represent any point of *complexified* Minkowski space-time M^* by a linear set like (2.10), but where not all the Z^α are null.[†]

We have, available, two alternative geometrical pictures for the description of phenomena (Fig. 1), namely the M-picture (which is the normal space-time description) and the C-picture, the space C being the three-complex-dimensional projective space of proportionally classes of twistors Z^α, of valence $\begin{bmatrix}1\\0\end{bmatrix}$. As a *real* manifold, the space C is six-dimensional. A five-real-dimensional submanifold N (topology $S^2 \times S^3$) of C defines the null twistors Z^α ($Z^\alpha \bar{Z}_\alpha = 0$). The removal of N from C leaves two disconnected open subsets C^+ and C^- (each with topology $S^2 \times E^4$) of C, defining, respectively, the right-handed ($Z^\alpha \bar{Z}_\alpha > 0$) and left-handed ($Z^\alpha \bar{Z}_\alpha < 0$) twistors Z^α. The manifold N may be thought of as the space of null lines in M (compactified, so M has topology $S^1 \times S^3$). We may refer to C^+ and C^- as spaces of 'complexified' null lines in M, but we must bear in mind that this 'complexification' process only increases the (real) dimensionality of the null line system from five to six. The points of M are represented in the C-picture as *complex projective lines* (topology S^2), which lie entirely within N. A projective line P, in C, which does not

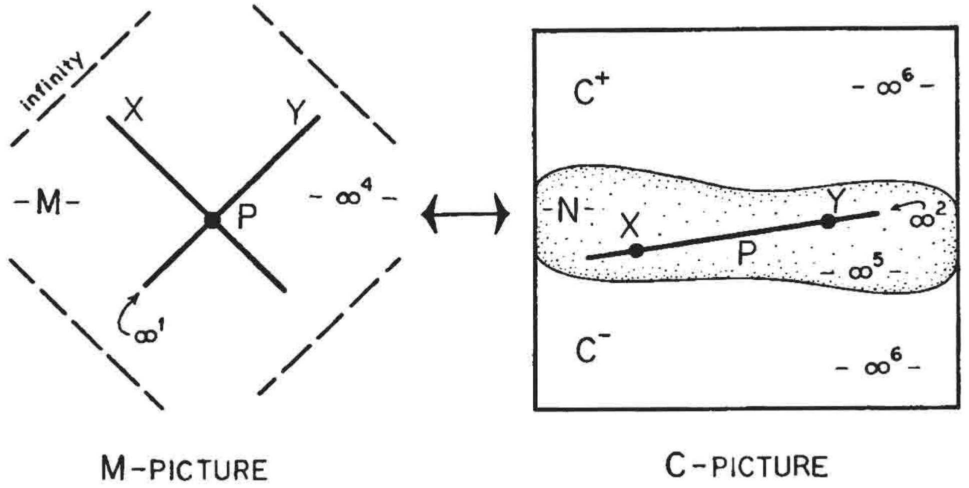


Figure 1.—The M-picture and C-picture representations of the twistors X^α, Y^α and $P^{\alpha\beta} = X^\alpha Y^\beta - Y^\alpha X^\beta$, where $X^\alpha \bar{X}_\alpha = Y^\alpha \bar{Y}_\alpha = 0 = X^\alpha \bar{Y}_\alpha$.

[†] In fact, M^* is just the *Grassmannian* (or Klein representation) of projective lines in the complex projective three-space C. (See, for example, Todd, J. A. (1947). *Projective and Analytical Geometry*. London; Semple, J. G. and Roth, L. (1949). *Introduction to Algebraic Geometry*. Oxford.

lie entirely on N represents a point of the *complexified* space M^*, for which the coordinates x^a are not all real. In fact, the *imaginary part* of x^a is spacelike, null, or timelike, respectively, according as P intersects N in a one-real-dimensional region (a curve: topology S^1), in a point, or not at all. If null or timelike, the imaginary part of x^a is future-pointing or past-pointing according as P lies in $C^+ \cup N$ or $C^- \cup N$.

A twistor W_α of valence $\begin{bmatrix}0\\1\end{bmatrix}$ defines a complex projective plane W in C (a four-real-dimensional submanifold of C), namely the set of points Z, of C, satisfying

$$W_\alpha Z^\alpha = 0 \tag{2.11}$$

Conversely, the plane W defines W_α uniquely up to proportionality. Now the plane W meets N in a three-real-dimensional set of points in the C-picture. These represent, in the M-picture a three-dimensional system of null lines (a null congruence). Such a congruence will define W uniquely. When W_α is null ($W_\alpha \overline{W}^\alpha = 0$), the W-congruence is simply the system of null lines in M meeting the null line $\overline{W}$ (with coordinates $\overline{W}^\alpha$). When W_α is right-handed ($W_\alpha \overline{W}^\alpha > 0$) [*resp.* left-handed ($W_\alpha \overline{W}^\alpha < 0$)], the W-congruence is a system of null lines in M, one through each point of M, which *twists* about every point in a *right-handed* [*resp. left-handed*] sense (a Robinson congruence). Thus, by invoking twistor complex conjugation, we can represent any twistor Z^α of valence $\begin{bmatrix}1\\0\end{bmatrix}$, up to proportionality, whether Z^α is null or not, by a (generally twisting) congruence of null lines in M, namely that defined by $W_\alpha = \bar{Z}_\alpha$.

We may think of twistor complex conjugation as defining a duality correspondence in the projective space C (a Hermitian correlation). To any point Z in C corresponds a unique 'polar' plane $\bar{Z}$; to any plane W in C corresponds a unique point $\overline{W}$ (the 'pole' of W). The point Z lies on the plane W if and only if $\overline{W}$ lies on $\bar{Z}$. The set N consists precisely of those points Z which lie on their 'polar' planes $\bar{Z}$. Dually, the plane W 'touches' N if and only if it contains its 'pole' $\overline{W}$. Similarly, a projective line P in C has a uniquely defined 'polar' line $\overline{P}$; Z lies on P if and only if $\overline{P}$ lies on $\bar{Z}$. The correspondence $P \leftrightarrow \overline{P}$ between lines in the C-picture represents, in the M-picture, precisely the correspondence between a point and its complex conjugate in the complexified space M^*. The *real* points of M^*, namely the points of M, are represented in the C-picture by the lines lying entirely on N. These are just the lines for which $P = \overline{P}$. More generally, the complex conjugation operation will apply to *any* twistor $K^{\alpha\beta\ldots\phi}_{\rho\ldots\tau}$ of general valence $\begin{bmatrix} p \\ q \end{bmatrix}$. The result is a twistor $\bar{K}^{\rho\ldots\tau}_{\alpha\beta\ldots\phi}$ of valence $\begin{bmatrix} q \\ p \end{bmatrix}$, where the labellings of the 0 and 1 components are interchanged with those of the 2 and 3 components on complex conjugation in the manner of (2.4); e.g. $\bar{K}^{2\ldots1}_{03\ldots2} = \overline{K^{21\ldots0}_{0\ldots3}}$ etc. In each case, any geometrical interpretation for $K^{\alpha\beta\ldots\phi}_{\rho\ldots\tau}$ in the C-picture will give rise to a corresponding *dual* interpretation for $\bar{K}^{\rho\ldots\tau}_{\alpha\beta\ldots\phi}$.

The allowable twistor transformations, other than those which involve a complex conjugate operation or a space or a time reversal in the M-picture, are given by

$$\tilde{K}^{\alpha\beta\ldots\phi}_{\rho\ldots\tau} = K^{\psi\zeta\ldots\theta}_{\lambda\ldots\nu}\, t_{\psi}{}^{\alpha}\, t_{\zeta}{}^{\beta} \ldots t_{\theta}{}^{\phi}\, \bar{t}_{\rho}{}^{\lambda} \ldots \bar{t}_{\tau}{}^{\nu} \tag{2.12}$$

where

$$t_{\beta}{}^{\alpha}\, \bar{t}_{\gamma}{}^{\beta} = \delta_{\gamma}{}^{\alpha} = \bar{t}_{\beta}{}^{\alpha}\, t_{\gamma}{}^{\beta}; \qquad |t_{\beta}{}^{\alpha}| = 1 \tag{2.13}$$

Because of the twistor complex-conjugation rule (here applied to $t_{\beta}{}^{\alpha}$), it follows that the group of $t_{\beta}{}^{\alpha}$-matrices satisfying (2.13) is just SU(2, 2). (The Hermitian form (2.5) has signature $(+,+,-,-)$.) In the C-picture, the transformations (2.12) (regarded as active transformations) are simply the projective transformations of C which leave N invariant and do not interchange C^{+} with C^{-}. In the M-picture these are the conformal transformations of M continuous with the identity.

Twistors can also be represented (a little more completely) in the M-picture in terms of certain spinor fields. If T_{α} is any twistor of valence $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$, we can define a 2-spinor field $\tau_{A}(x^{a})$ by

$$\tau_{A} = \kappa_{A} - i x_{AB'} \rho^{B'} \tag{2.14}$$

where $T_0 = \kappa_0$, $T_1 = \kappa_1$, $T_2 = \rho^{0'}$, $T_3 = \rho^{1'}$. Then τ_A transforms correctly as a conformal density of weight $\frac{1}{2}$ as T_{α} undergoes (2.12). In fact, a spinor field of the type given by (2.14) (where κ_A and $\rho^{B'}$ are constant) is the general solution of the equation

$$\nabla_{P'(A}\, \tau_{B)} = 0 \tag{2.15}$$

where round brackets denote symmetrisation and where

$$\nabla_{P'A} \equiv \partial/\partial x^{AP'} \tag{2.16}$$

(so, equivalently, $\nabla^{P'A} \equiv \partial/\partial x_{AP'}$). The field τ_A gives a conformally invariant M-picture realisation of the twistor T_{α} up to a multiple of ± 1, $\pm i$. (This fourfold ambiguity arises because the spinor field τ_A picks up a factor $\pm i$ as it crosses the null cone at infinity. Such an ambiguity is essential because of the four-valued nature of the $\begin{bmatrix}0\\1\end{bmatrix}$-twistor representation of the conformal group.) Any completely symmetric covariant twistor $S_{\alpha\beta\ldots\phi}=S_{(\alpha\beta\ldots\phi)}$ also has an M-picture realization as a spinor field somewhat similar to (2.14). For example, for valence $\begin{bmatrix}0\\2\end{bmatrix}$, if (for $A,B=0,1$) we put $S_{AB}=\lambda_{AB}$, $S_{A,B+2}=\mu_A^{B'}$, $S_{A+2,B+2}=\nu^{A'B'}$ (where $S_{\alpha\beta}=S_{\beta\alpha}$ so $\lambda_{AB}=\lambda_{BA}$, $\nu^{A'B'}=\nu^{B'A'}$), the spinor field

$$\sigma_{AB}=\lambda_{BA}-2i\mu^{C'}_{(A}x_{B)C'}-\nu^{C'D'}x_{AC'}x_{BD'} \tag{2.17}$$

is of the type of the general symmetric solution of

$$\nabla_{P'(A}\sigma_{BC)}=0 \tag{2.18}$$

and represents $S_{\alpha\beta}$ in a conformally invariant way up to sign ± 1. A general (non-symmetric) twistor of valence $\begin{bmatrix}p\\q\end{bmatrix}$ has a representation similar to (2.17), but only as a many-point field.

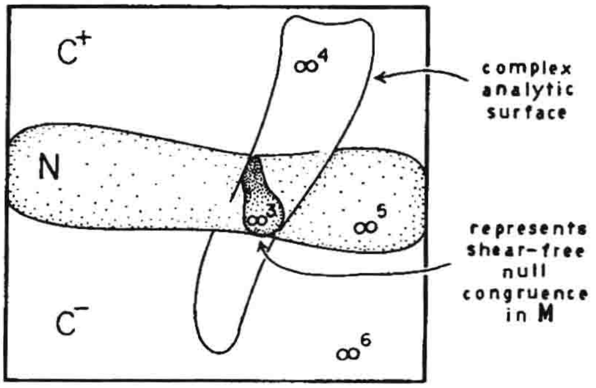


Figure 2.—The Kerr theorem.

The geometrical significance to the M-picture of the analytic structure of the C-picture is best illustrated by the *theorem of Kerr* (Kerr, unpublished). Let Q be a complex analytic surface in C, so Q is defined by the vanishing of a homogeneous analytic function in the twistor coordinates Z^α. The set Q is four-real-dimensional and intersects N in a three-real-dimensional region (Fig. 2). This defines a congruence of null lines in M. Kerr's theorem states that such a congruence is necessarily *shear-free* and conversely that any shear-free congruence of null lines in M is obtainable in this way (or as a limiting case of such a construction). The shear-free condition states that if a 'bundle' of null lines of the congruence neighbouring a given null line X, and lying (to first order) in the null hyperplane through X, has a (small) circular cross-section at one point of X, then it also has a (small) circular cross-section at every other point of X (Fig. 3). (The Robinson congruences, given when Q is a plane, are particular examples with this property.) The shear-free condition appears as a kind of M-picture realisation of the Cauchy-Riemann equations for C.

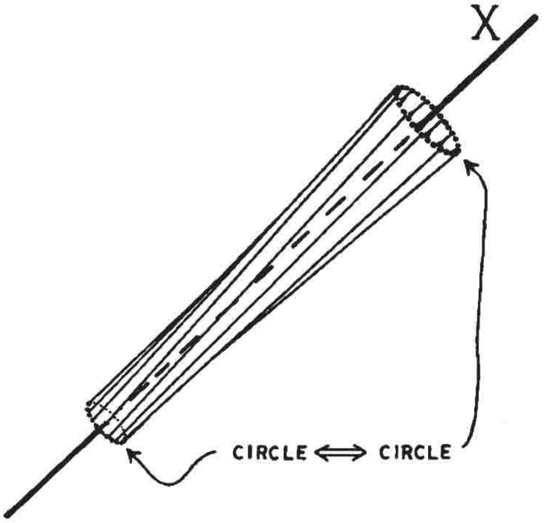


Figure 3.—The shear-free condition. (The circles lie, to first order, in the null hyperplane containing X.)

3. *Twistor Description of Zero Rest-Mass Fields*

In another paper (Penrose, 1968), the following contour integral formula was introduced, which expresses the general (analytic) solution of the spin s, zero rest-mass, free-field equations, for

$$s = 0, \tfrac{1}{2}, 1, \tfrac{3}{2}, \ldots$$

in terms of an arbitrary analytic function of three complex variables $F(\lambda, \mu, \nu)$:

$$\phi_r = \frac{1}{2\pi i} \oint \lambda^r F(\lambda, u + \lambda\bar{\zeta}, \zeta + \lambda v)\, d\lambda \quad (r = 0, 1, \ldots, 2s) \tag{3.1}$$

(The contour is to surround singularities of F and to vary continuously with u, v, ζ.) For, we have $\partial\phi_r/\partial\bar{\zeta} = \partial\phi_{r+1}/\partial u$ and $\partial\phi_r/\partial v = \partial\phi_{r+1}/\partial\zeta$ $(r = 0, \ldots, 2s - 1)$, if $s > 0$, and $\{\partial^2/\partial u\,\partial v - \partial^2/\partial\zeta\,\partial\bar{\zeta}\}\phi_0 = 0$ for $s = 0$. Writing

$$\phi_0 = \phi_{00\ldots0}, \qquad \phi_1 = \phi_{10\ldots0}, \qquad \ldots, \qquad \phi_{2s} = \phi_{11\ldots1} \tag{3.2}$$

these equations become, if $s > 0$,

$$\nabla^{P'A} \phi_{AB\ldots L} = 0 \tag{3.3}$$

where $\phi_{AB\ldots L}$ is symmetric $(= \phi_{(AB\ldots L)})$ with $2s$ indices and

$$\nabla_{P'A} \nabla^{P'A} \phi = 0 \tag{3.4}$$

if $s = 0$. Equations (3.3) and (3.4) are simply the zero rest-mass free-field equations (Dirac, 1936; Fierz, 1940) for spin $s > 0$ (Dirac-Fierz) and spin $s = 0$ (D'Alembert), respectively.

Thus the free-field equations on $\phi_{AB\ldots L}$ are an automatic consequence of the analyticity of F. The converse result that *any* analytic zero rest-mass free field $\phi_{AB\ldots L}$ has a representation in the form (3.1) is also true, but the argument giving the construction of F from $\phi_{A\ldots L}$ will not be entered into here. We note, in this context, that the function F is not, however, uniquely determined by $\phi_{A\ldots L}$. If, for example, we add to F any function which is regular inside the contour (for every u, v, ζ), then clearly the resulting field $\phi_{A\ldots L}$ will be unaffected. This freedom of choice for F defines a kind of gauge group (different from the usual gauge group) but which depends on the positioning of the contour (as a function of u, v, ζ). The nature of this gauge freedom will not be discussed here. Nor will certain interesting features of the representation (3.1), such as the fact that 'null' or 'algebraically special' solutions of (3.3) can be very readily generated, by merely requiring that the contour surround only a simple or low-order pole of F (Penrose, 1968). Instead, two matters of more immediate relevance will be treated, namely the transcription of (3.1) into a more general twistor form and a topological requirement on the singularities of F that ensures that the field $\phi_{A\ldots L}$ has positive (or, alternatively negative) frequency. The question of a conformally invariant Hilbert space norm for fields $\phi_{A\ldots L}$, given in terms of F, will be discussed in Section 5.

We note, first, that using the notation (2.6) or (2.8) we can write

$$F(\lambda, u + \lambda\bar{\zeta}, \zeta + \lambda v) \equiv F\left(-\frac{Z^0}{Z^1}, \frac{iZ^3}{Z^1}, -\frac{iZ^2}{Z^1}\right) \tag{3.5}$$

where

$$du + \lambda\, d\bar{\zeta} = 0 = d\zeta + \lambda\, dv \tag{3.6}$$

Set

$$f(Z^\alpha) \equiv f(Z^0, Z^1, Z^2, Z^3) \equiv (iZ^1)^{-2s-2} F\left(-\frac{Z^0}{Z^1}, \frac{iZ^3}{Z^1}, -\frac{iZ^2}{Z^1}\right) \tag{3.7}$$

Then $F(Z^\alpha)$ is analytic and *homogeneous* of degree $-2s-2$ in Z^α. The twistor Z^α represents a null line Z through the point P with coordinates u, v, ζ. The direction of Z is defined by λ according to (3.6). Let X and Y be the two particular null lines through P given respectively by $\lambda=\infty$ and $\lambda=0$. We can assign the twistors X^α, Y^α to X, Y as follows:

$$(X^\alpha)=(i,0,v,-\bar{\zeta});\qquad (Y^\alpha)=(0,-i,\zeta,-u) \tag{3.8}$$

and set

$$Z^\alpha=\lambda X^\alpha+Y^\alpha \tag{3.9}$$

This is consistent with (2.6) and (3.6), and we have $Z^1=-i$. By (3.7) and (3.5), we can now write the formula (3.1) as

$$\phi_r(X^\alpha,Y^\alpha)=\frac{1}{2\pi i}\oint\lambda^r f(\lambda X^\alpha+Y^\alpha)\,d\lambda\quad(r=0,1,\ldots,2s) \tag{3.10}$$

The quantities ϕ_r are here written explicitly as functions of X^α and Y^α. This is because the spinor components (3.2) are dependent on a particular choice of spin frame, and although this spin frame was originally related to the coordinate system (2.1), we can think of X^α and Y^α, instead, as defining the spin frame at P. Indeed, we can now dispense with our original coordinates (2.1) altogether, since the point P is defined by X^α and Y^α via

$$P^{\alpha\beta}=X^\alpha Y^\beta-Y^\alpha X^\beta \tag{3.11}$$

which serve as the twistor coordinates for P. We can, in fact, enlarge the domain of the formula (3.10) by dropping the requirement that X^α and Y^α have the special form (3.8), retaining merely (for a real point P) the condition that X^α and Y^α both be null and satisfy $X^\alpha\bar{Y}_\alpha=0$. Keeping the $P^{\alpha\beta}$ of (3.11) fixed, the freedom of choice for X^α and Y^α is given by the 'spin transformation':

$$\hat{X}^\alpha=\rho X^\alpha+\sigma Y^\alpha,\qquad \hat{Y}^\alpha=\tau X^\alpha+\omega Y^\alpha \tag{3.12}$$

with

$$\rho\omega-\sigma\tau=1 \tag{3.13}$$

where inversely

$$X^\alpha=\omega\hat{X}^\alpha-\sigma\hat{Y}^\alpha,\qquad Y^\alpha=-\tau\hat{X}^\alpha+\rho\hat{Y}^\alpha \tag{3.14}$$

Substituting (3.14) into (3.10), putting

$$\lambda=(\rho\mu+\tau)/(\sigma\mu+\omega) \tag{3.15}$$

and keeping the contour 'fixed' (i.e. so that the same values of $f(Z^\alpha)$ are involved after the substitution), we get

$$\phi_r(X^\alpha,Y^\alpha)=\frac{1}{2\pi i}\oint(\rho\mu+\tau)^r(\sigma\mu+\omega)^{2s-r}f(\mu\hat{X}^\alpha+\hat{Y}^\alpha)\,d\mu \tag{3.16}$$

When expressed in terms of

$$\phi_r(\hat{X}^\alpha, \hat{Y}^\alpha) = \frac{1}{2\pi i}\oint \mu^r f(\mu\hat{X}^\alpha + \hat{Y}^\alpha)\, d\mu \quad (r = 0, 1, \ldots, 2s) \quad (3.17)$$

(3.16) yields the transformation law for a $D(s, 0)$ representation† of the homogeneous Lorentz group at P. This is precisely what is required of the formula (3.10), in order that it should correspond to a field $\phi_{A\ldots L}$ correctly transforming under local Lorentz transformations. It is for this purpose that the function f, as defined in (3.7), is chosen to be *homogeneous of degree* $-2s - 2$. With the original expression (3.1), the role played by the value of s does not emerge. The full twistor expression (3.10), on the other hand, contains the *entire transformation behaviour* of the field.

We may note the general twistor form of the equations (3.3):

$$\frac{\partial\phi_r}{\partial X^\alpha} = \frac{\partial\phi_{r+1}}{\partial Y^\alpha} \quad (r = 0, 1, \ldots, 2s - 1) \quad (3.18)$$

and, for $s = 0$, of (3.4):

$$\frac{\partial^2\phi_0}{\partial X^\alpha \partial Y^\beta} = \frac{\partial^2\phi_0}{\partial Y^\alpha \partial X^\beta} \quad (3.19)$$

(The twistor 'wave equation' (3.19) applies also, of course, if $s > 0$ with ϕ_0 replaced by any of the ϕ_r.) Each ϕ_r is separately homogeneous of degree $-(r + 1)$ in X^α and degree $-(2s - r + 1)$ in Y^α, so that by Euler's theorem

$$X^\alpha \frac{\partial\phi_r}{\partial X^\alpha} = -(r + 1)\phi_r, \qquad Y^\alpha \frac{\partial\phi_r}{\partial Y^\alpha} = -(2s - r + 1)\phi_r \quad (3.20)$$

Thus (3.18) or (3.19), together with (3.20), are the *twistor versions of the zero rest-mass free-field equations*. Note that in twistor space, a field (with spin) is regarded as a *two-point function*. The *line P* in the C-picture which connects the *two points* X and Y (of C) corresponds, in the M-picture, to the *point* P at which the field $\phi_{A\ldots L}$ is evaluated, this being the intersection of the two *null lines* X and Y. The dependence of the ϕ_r on the positioning of the points X, Y on P (C-picture) corresponds to the *spin structure* of the field (this being given, in the M-picture, by the dependence of $\phi_{A\ldots L}$ on choice of spin frame at the point P—as defined by X^α, Y^α). Note that if $s = 0$, then ϕ is actually independent of X^α and Y^α except in so far as they define $P^{\alpha\beta}$

---

† See, for example, Ya Liubarski, G. (1960). *The Application of Group Theory in Physics* (trans. Dedijer, S.). Pergamon Press, New York; or Streater, R. F. and Wightman, A. S. (1964). *PCT, Spin and Statistics and All That*. W. A. Benjamin, New York.

according to (3.11). (This follows† from $X^\alpha \partial\phi_0/\partial Y^\alpha = 0 = Y^\alpha \partial\phi_0/\partial X^\alpha$, which is a consequence of (3.19) and (3.20).) This simply reflects the scalar nature of the field in this case.

The formula (3.10) can also be applied when X^α and Y^α are not null and need not satisfy $X^\alpha \bar{Y}_\alpha = 0$. Then the line P of the C-picture, which connects the two points X and Y [$P^{\alpha\beta}$ given as in (3.11)], does not lie on N and so represents a *complex* point $P \in M^*$ in the M-picture. The M-picture realisation of (3.10) in its full generality is thus a *complexified* zero rest-mass field. It is this fact that enables us to give a direct C-picture interpretation of a condition which ensures that $\phi_{A\ldots L}$ is a positive (or alternatively negative) frequency field. (We recall that *any* C-picture property has an interpretation in terms of the *real* space M. Thus, even the complex points of M^* can be interpreted in terms of real structures (Penrose, 1967a) in M. It is interesting to note, in this context, that much of the structure of $f(Z^\alpha)$ has a very *direct* 'real' interpretation in the M-picture. For *null* Z^α we may think of $f(Z^\alpha)$ as defining a function on the space of null lines of M, although the dependence of f on the scaling of Z^α must be borne in mind. This function enters directly into defining the field $\phi_{A\ldots L}$ at real points P and also at some complex points P.)

Let us consider the nature of the condition of 'positive frequency' as applied to a field $\phi_{A\ldots L}$. I shall be concerned only with fields which are non-singular throughout the entire (conpactified) space-time M. This will imply, in particular, that the field enjoys suitable asymptotic behaviour (i.e. in the neighbourhood of the null cone at infinity), yielding the 'peeling-off' property for the field (Penrose, 1965). In this context we must take into account the phenomenon of Grgin (1966): for a free-wave field we must expect the field to jump by a factor $(\pm i)^{2s+2}$ as we cross the null cone at infinity. Owing to the four-valued nature of the M-picture realisation of twistors, this is indeed the type of behaviour that we should expect from (3.10). I shall, in fact, regard the field $\phi_{A\ldots L}$ as being appropriately 'analytic across infinity' *only* if it possesses the correct Grgin discontinuity there. Then the fields under consideration will be imploding-exploding free-waves which tail off suitably towards infinity in all directions. In twistor terms, this condition is simply that $\phi_r(X^\alpha, Y^\alpha)$ should be regular for all null X^α, Y^α for which $X^\alpha \bar{Y}_\alpha = 0$.

We can imagine the field $\phi_{A\ldots L}$ decomposed into plane waves. The various amplitudes of these waves will be given by the *null-datum* (Penrose, 1963, 1967b) for $\phi_{A\ldots L}$ at all the various points on the null cone at infinity. The null-datum for $\phi_{A\ldots L}$ at a point P, for a null hypersurface containing a null line X through P, is simply

---

† See, for example, Hodge, W. D. and Pedoe, D. (1952). *Methods of Algebraic Geometry*, Vol. II, Cambridge University Press.

$$\phi \equiv \phi(X^\alpha; P^{\alpha\beta}) \equiv \phi_0(X^\alpha, Y^\alpha) \tag{3.21}$$

where $P^{\alpha\beta}$ is related to X^α, Y^α by (3.11). (The fact that $\phi_0(X^\alpha, Y^\alpha)$ depends on X^α and $P^{\alpha\beta}$, and not on Y^α, is a consequence of (3.14) and (3.16), with $\sigma = 0$, $\omega = 1$.) The amplitude for the plane-wave component of $\phi_{A\ldots L}$ corresponding to a *fixed* null direction is given by (3.21) where X^α is kept fixed and $P^{\alpha\beta}$ is allowed to vary. Here X is the generator of the null cone at infinity which corresponds to this fixed null direction (Fig. 4); P is a variable point on X which defines a *null hyperplane of constant amplitude* for the wave. (P is the 'vertex' of this null hyperplane.)

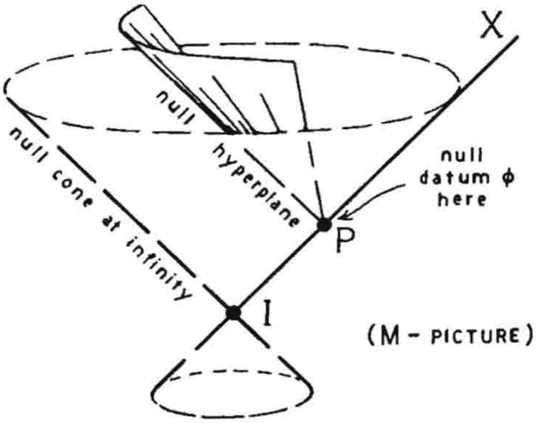


Figure 4.—A null hyperplane of constant amplitude, in a plane wave, has a vertex P on the null cone at infinity. The amplitude is measured by the complex null datum $\phi(X^\alpha; P^{\alpha\beta})$.

We must express the condition that the complex quantity ϕ, as a function of the real variable defining the position of P on the line X, has positive frequency. Let us imagine† X as a finite line. (The concept of a 'positive frequency function' defined on a null line is, in fact, invariant under restricted conformal transformations of M, since these give the restricted projective group on a real null line.‡

---

† To be more precise, 'imagine ...' can be interpreted to mean 'make a conformal transformation on M, continuous with the identity, so that after the transformation we have ...'. Thus, we are at liberty to choose coordinates so that certain points at infinity receive finite coordinates.

‡ See Gel'fand, I. M., Graev, M. I. and Vilenkin, N. Ya. (1966). *Generalized Functions V: Integral Geometry and Representation Theory*. Academic Press, New York and London.

But this is not really needed here, since we can imagine I to remain at infinity.)

For simplicity, imagine our origin of coordinates to be situated on X. Then the real points of X have position vectors of the form ul^a, where u is a real number and l^a is a real future-pointing null vector. Along X, we now have ϕ as a complex function of the real variable u. This defines the amplitude of our plane-wave component. The wave has positive frequency if this function can be extended to a complex analytic function of the *complex* variable $u + iw$ which is regular for $w > 0$. (The regular behaviour at infinity of ϕ will be taken care of by the fact that we are really concerned with entire *compact* manifold M.) Now, the position vectors $(u + iw)l^a$, with $w > 0$, have null future-pointing imaginary parts. Therefore (cf. Section 2) they define points P of M^* which correspond in the C-picture, to lines lying in $N \cup C^+$. If $\phi(X^\alpha; P^{\alpha\beta})$ remains regular for these points, then we shall have the required analytically extended function. (The fact that ϕ is a *complex analytic* function of $u + iw$ follows from the general analytic properties of the C-picture, but this can also be seen explicitly if we go back to the definition of ϕ_0 given by (3.1), where u is allowed to be replaced by the complex variable $u + iw$). Thus, if we can construct $\phi_r(X^\alpha, Y^\alpha)$ which is regular *whenever* the line joining X and Y lies in $N \cup C^+$, we shall certainly have constructed a field of positive frequency. (In fact *all* fields of positive frequency will have this apparently more general property, but the matter will not be entered into here.)

Now let us consider a condition on f which ensures that ϕ_r is of this type. We suppose, in fact, that the points Z in $N \cup C^+$, for which $f(Z^\alpha)$ is singular, can be contained in *two disjoint closed proper subsets* S_+, S_- of $N \cup C^+$. For each line P, in C, which intersects both of $S_\pm$, we shall choose our contour so as to surround the whole of $P \cap S_+$ once in a positive sense (on the Riemann sphere S^2 which represents the points of P). Thus, the contour correspondingly surrounds the whole of $P \cap S_-$ once in a negative sense. (This can always be arranged since $P \cap S_+$ and $P \cap S_-$ are disjoint closed sets. The contour need not be a connected curve but may consist of several loops.) If either of S_+ or S_- is vacuous the situation becomes trivial and the field ϕ_r would have to be zero. I shall suppose that ϕ_r is, in fact, not identically zero. Then at least some lines in C must intersect both of $S_\pm$. In fact, if $P \subset N \cup C^+$, then P *necessarily* intersects both of S_+, S_-, since otherwise every line in $N \cup C^+$ 'sufficiently close' to P would share P's behaviour and we should have an open set in M^* on which ϕ_r vanishes. This would imply $\phi_r \equiv 0$, since ϕ_r is analytic, contrary to hypothesis.

We see that the field ϕ_r is uniquely determined, for $P \subset N \cup C^+$

(and for certain other P's also) by the division of the singularities of f between the two sets S_+, S_-. No more precise specification of the contour is required. Thus, by the preceding discussion, $\phi_r(X^\alpha, Y^\alpha)$ will be defined and analytic whenever the join of X and Y lies in $N \cup C^+$, whence ϕ_r is necessarily a field of *positive frequency*. To obtain a field of negative frequency we need only apply the identical construction but with C^- replacing C^+. We can also generalise the discussion, somewhat, if we consider functions $f(Z^\alpha)$ where the singularities need be disconnected only in C^+ rather than in $N \cup C^+$. The resulting fields ϕ_r would then not necessarily be regular at all real points P of M. This would enable us to consider non-analytic fields and distributions (Green's functions) on M as boundary values of complex analytic fields.†

It is natural to ask how general the above construction is, for positive frequency fields. It appears, in fact, to be as general as one would wish it to be, but a complete argument is at the moment lacking. To indicate something of the generality involved, however, I can exhibit some examples. Let Q_α and R_α be two right-handed twistors for which every linear combination is also right-handed. This is achieved by taking $\bar{Q}$ and $\bar{R}$ as two points of a line lying entirely in C^+ and we have

$$Q_\alpha \bar{Q}^\alpha > 0, \qquad R_\alpha \bar{R}^\alpha > 0, \qquad Q_\alpha \bar{Q}^\alpha R_\beta \bar{R}^\beta > |Q_\alpha \bar{R}^\alpha|^2 \tag{3.22}$$

The intersection of the two planes Q and R (being the 'polar line' of the line joining $\bar{Q}$, $\bar{R}$) will lie entirely in C^-. Thus $S_+ = Q \cap (N \cup C^+)$ and $S_- = R \cap (N \cup C^+)$ will be disjoint closed subsets of $N \cup C^+$ containing the relevant singularities of

$$f(Z^\alpha) \equiv (Q_\alpha Z^\alpha)^{-1}(R_\beta Z^\beta)^{-2s-1} \tag{3.23}$$

Of course (if $s > 0$) we could allow other negative powers of $Q_\alpha Z^\alpha$. Furthermore, we can form various finite linear combinations of expressions like (3.23), with differing Q's and R's. We could also take combinations of expressions like (3.23), but with polynomials in Z^α in the numerator and higher powers in the denominator. These would all yield f's which were algebraic functions of Z^α, but we could generate non-algebraic functions also satisfying the required conditions, simply by integrating together expressions like (3.23) over some (not too extensive) domain. That this yields positive frequency fields ϕ_r of considerable generality, at least, follows from the fact that in the limit when the inequalities (3.22) tend to equalities, the field generated by (3.23) becomes the 'Feynman propagator' for $\phi_{AB\ldots L}$ appropriate to the characteristic initial value problem.

---

† See, for example, Streater, R. F. and Wightman, A. S. (1964). *PCT, Spin and Statistics and All That*. W. A. Benjamin, New York.

Finally, it may be mentioned that there is an interesting topological property of such 'positive frequency' singularity sets in relation to the topology of N, as imbedded in C. I shall illustrate this explicitly, only in the case of f given by (3.23). As was mentioned in Section 2, we have $N \cong S^2 \times S^3$, $C^- \cong C^+ \cong S^2 \times E^4$. We may regard C^+ as being obtained from N by 'filling in' each S^3 with a cell E^4. The complex plane Q meets N in an S^3 of the product $S^2 \times S^3$. The intersection of Q with C^+ is an E^4 of the product $S^2 \times E^4$ which 'fills in' this S^3. Similarly, the complex plane R meets N in another S^3 of the product $S^2 \times S^3$ which is again 'filled in' by an E^4 $(= R \cap C^+)$ of the product $S^2 \times E^4$ $(= C^+)$. In fact the *pencil of planes* defined by Q and R (with twistors $Q^\alpha + \mu R^\alpha$) generates the entire 'filling in' of $S^2 \times S^3 = N$ by $S^2 \times E^4 = C^+$. Now, from the symmetry between C^+ and C^-, it follows that C^- must also 'fill in' N in exactly the same way. However, this new 'filling in' applies to an *essentially different way of expressing N as $S^2 \times S^3$* from the one given above. If we consider N as imbedded in $N \cup C^-$ rather than in $N \cup C^+$, it emerges that the sets $Q \cap N$ and $R \cap N$ are *linked* in the sense that any submanifold of C^- which spans $Q \cap N$ (i.e. whose boundary is $Q \cap N$) *must necessarily* intersect any submanifold of C^- which spans $R \cap N$. (Furthermore, neither $Q \cap N$ nor $R \cap N$ spans an E^4 in C^-.) Thus, we see that a particular topological structure of the singularity regions of our 'positive frequency' function $f(Z^\alpha)$ emerges even in the region $Z \in N$. The regions $N \cap S_+$ and $N \cap S_-$ twist around one another in a way opposite to the corresponding behaviour for a 'negative frequency' function. This is realised, in the M-picture, as a right-handed twist (for 'positive frequency' f) and a left-handed twist (for 'negative frequency' f) for the null line systems concerned.†

---

† The reader may be disturbed by this association of a screw-sense in space-time with the notion of 'positive frequency' for a zero rest-mass field. In the present formalism, the opposite association would have been achieved had we been concerned with a spinor field $\psi_{A'B'\ldots L'}$ rather than $\phi_{AB\ldots L}$. With half-odd spin fields this association of a particle helicity with the field is familiar. Here, this must be applied also to integral spin. Thus ϕ_{AB} describes the 'photino' (or positive helicity photon) while $\bar{\phi}_{A'B'}$ describes the 'anti-photino'. Interactions will readily convert 'photinos' into 'anti-photinos'.

The fact that the use of twistors entails an interrelation between '$+i$', 'right-handedness', and 'future-orientation' will be evident from the interpretation of the *lines* in C^+ as points of M^* with position vectors having future-pointing imaginary parts, and the interpretation of *points* of C^+ in terms of right-handed Robinson congruences.

4. *Twistors for Curved Space-time*

We have seen, in Section 3, the importance of the complex analytic structure of the C-picture in the treatment of zero rest-mass fields. On the other hand, it is (apparently) precisely this complex analytic structure which is destroyed by the presence of conformal curvature (i.e., gravitation) in the M-picture. For Kerr's theorem tells us that the Cauchy-Riemann equations for the C-picture have an interpretation as a 'shear-free' condition on null line systems in the M-picture. But when a shear-free 'bundle' of null lines (geodesics) enters a region of conformal curvature in space-time it will emerge on the other side *possessing* shear, in general. Thus, if M is conformally curved, we cannot interpret 'shear-freeness' as referring to null lines in their entirety, but only to null lines 'in the neighbourhood of a point'. We *may* think of a C-picture as referring accurately only to a 'sufficiently small' neighbourhood of a point of M, for which the conformal curvature can be neglected. But we would like, also, to be able to discuss the interconnections between different neighbourhoods, so that the real effect of conformal curvature (that is, of gravitation) on C-picture structure can be investigated. It will emerge, in fact, that what is required is not a different C-picture for each 'neighbourhood of a point' in space-time, but rather one C-picture for the whole of space-time (roughly speaking) whose 'complex analytic structure' appears to 'shift' as we move about the space-time, and which possess a (weaker) *symplectic* structure which does *not* shift. This will be the *classical* C-picture for M which we consider in this section. The passage to a quantised theory (Section 5) which this classical C-picture structure suggests will, in a certain sense, reinstate the full C-picture analyticity.

Let us suppose that M is a four-dimensional manifold with a pseudo-Riemannian metric $ds^2 = g_{ab}dx^a dx^b$, of signature $(+, -, -, -)$, and which possesses suitable global properties. These global restrictions on M will not concern us particularly here and we may, if desired, simply restrict our attention to some suitably well-behaved open submanifold of space-time. The idea will be to represent the null geodesics of M as the points of some five-dimensional manifold N. The *invariant structure* of N will be defined in terms of properties of null geodesics in M which do not refer to *particular* points on these null geodesics, but which can be read off by examining the geodesics at *any* of their points. For example, the property of a system of null geodesics that they should generate a *null hypersurface*,† will turn out to represent an aspect of the invariant structure of N. On the other hand, the property of two null geodesics that they *intersect*, will not. It will emerge that the invariant structure of N will be most easily describable if we regard N as a submanifold of a six-real-dimensional manifold C. In this sense, the classical C-picture will have value for the description even of conformally curved space-time.

---

† That is, so that the tangent 3-space to the hypersurface is tangent to the light cone. This property 'propagates' in the sense that it holds globally (except for singularities) if it holds on any space-like cross-section of the hypersurface.

The discussion will be given here in terms of certain idealised space-times in the first instance, namely plane waves—or, more generally, plane-fronted waves (Brinkmann, 1923; Robinson, 1958)—for which the amplitudes can be given by Dirac delta functions. The idea will then be to regard the effect on the C-picture, of a general region of M-picture curvature, as a (non-linear) composition of effects of such plane-fronted waves.

The general plane-fronted wave has a metric which can be put into the form

$$ds^2 = 2(du + R(v, \zeta, \bar{\zeta})\, dv)\, dv - 2d\zeta\, d\bar{\zeta} \tag{4.1}$$

(R being real). The non-vanishing curvature tensor components for the metric (4.1) are defined by

$$\partial^2 R/\partial\zeta^2, \qquad \partial^2 R/\partial\zeta\,\partial\bar{\zeta}, \qquad \partial^2 R/\partial\bar{\zeta}^2 \tag{4.2}$$

The metric satisfies Einstein's vacuum equations (and so represents a plane-fronted purely gravitational wave) if

$$\partial^2 R/\partial\zeta\,\partial\bar{\zeta} = 0 \tag{4.3}$$

but we shall not be concerned with implications of (4.3) here. If (4.3) is *not* imposed, then the metric (4.1) covers the more general situation of a combined plane-fronted gravitational-electromagnetic-neutrino wave.

If $R = 0$ over some range of v, then we have a region of Minkowski space-time [compare (2.3)]. If R is non-zero only within some finite range of v: $v_1 < v < v_2$, then we have the situation of a 'sandwich wave'. In the idealized situation where the wave is allowed to become infinitesimal in duration (say, $v_1, v_2 \to 0$) while still producing a non-zero resultant effect, the function R becomes a delta function in v:

$$R(v, \zeta, \bar{\zeta}) \equiv \delta(v)\, r(\zeta, \bar{\zeta}) \tag{4.4}$$

With the substitution (4.4), the form of metric (4.1) does not satisfy the conditions normally required (Bruhat, 1959) for a space-time with delta functions in the curvature, since we here have a delta function in the *metric* tensor components. Under normal circumstances this would

CHAPTER 37

lead to non-allowable *products* of delta functions in addition to derivatives of delta functions in the curvature. However, in the present situation, the space-time represented by (4.1), (4.4) is still allowable. It is, in fact, possible to make a coordinate substitution so that the new metric tensor components are C^0 functions of the coordinates. Then the fact that this leads only to a simple delta function (C^{-2}) type of curvature becomes more obvious. But for our purposes the form (4.1) is more satisfactory since it leaves the flat portions of the space-time in the required Minkowski form (2.3). The (delta function) curvature components are then obtained from the substitution of (4.4) into (4.2).

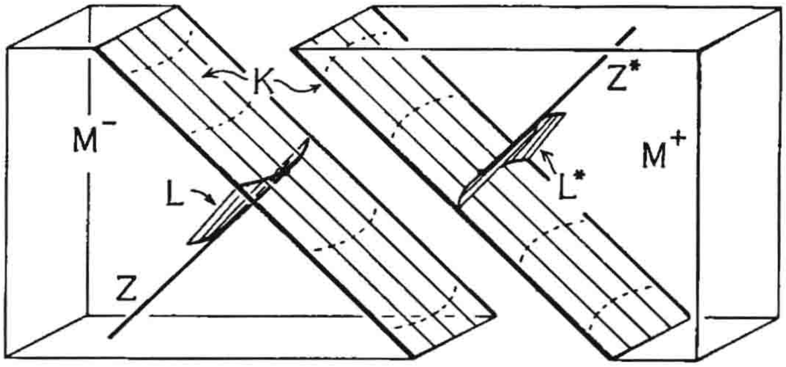


Figure 5.—The two Minkowski half-spaces are joined in a 'warped' fashion along the null hyperplane K. The continuation Z^* of the null line Z can be defined in terms of the null hypersurface L, which continues as L^*.

We can describe the resulting manifold in the following 'scissors and paste' terms. We divide ordinary Minkowski space-time (metric $ds^2 = 2du\,dv - 2d\zeta d\bar{\zeta}$) into two portions M^-, M^+ by the removal of the null hyperplane $v = 0$. Thus M^- is given by the portion $v < 0$ and M^+ by $v > 0$. We wish to join M^- and M^+ together again, inserting a null hyperplane K as their common boundary, but in such a way that the imbedding of K in each of $M^\pm$ appears 'warped' as viewed from the other (Fig. 5). More specifically, $M^- \cup K$ and $M^+ \cup K$ each have the normal Minkowski metric (2.3), but the two halves are joined in a way not consistent with a four-dimensional Minkowskian metric structure at K. The three-dimensional metric induced on K by its imbedding in each of the two halves is the same, however. (A somewhat analogous two-dimensional example is obtained if we imagine two ordinary cones joined base to base. The two surfaces are intrinsically flat, but a delta function in the curvature resides along the edge at which the cones are joined.) For convenience, we use coordinates u, v, ζ in M^- and u^*, v^*, ζ^* in M^+, the entire manifold M being defined by:

$$M^- \cup K: \quad ds^2 = 2du\,dv - 2d\zeta\,d\bar{\zeta}$$

$$(v \leqslant 0 \text{ with } u,\ \zeta \text{ unrestricted; } u \text{ real})$$

$$M^+ \cup K: \quad ds^2 = 2du^*\,dv^* - 2d\zeta^*\,d\bar{\zeta}^* \tag{4.5}$$

$$(v^* \geqslant 0 \text{ with } u^*,\ \zeta^* \text{ unrestricted; } u^* \text{ real})$$

$$K: \quad v^* = v = 0, \qquad \zeta^* = \zeta, \qquad u^* = u - r(\zeta, \bar{\zeta})$$

Thus, the generators of K are 'shunted down' by an amount $r(\zeta, \bar{\zeta})$ when we pass from M^- to M^+. Equations (4.5) are just another way of expressing what is meant by the metric (4.1) when the substitution (4.4) is made.

The manifold M possesses two regions M^- and M^+ which are exactly flat. Thus, we can construct a C-picture in terms of either of these regions. A null line in M^- can be given twistor coordinates Z^α according to the scheme (2.8); similarly a null line in M^+ can be given twistor coordinates $Z^{*\alpha}$ according to the 'starred' version of (2.8). But any null line in M^- which intersects K (i.e., which is not parallel to the null direction in K) will emerge as the *uniquely defined* null line, in M^+, for which the two portions constitute a 'null geodesic' in M. We may think of such a 'null geodesic' as resulting when a limiting process is applied to null geodesics for spaces (4.1), as the function $R(v, \zeta, \bar{\zeta})$ approaches the form (4.4). More conveniently, there is, however, a direct geometrical construction of these 'null geodesics'. This arises owing to the fact that if null geodesics generate a null hypersurface in one portion of a space-time manifold, then they must continue to generate a null hypersurface in any other portion of the manifold. We must *retain* this property for the 'null geodesics' in M (as follows from the above limit construction). Thus, if the null line Z in $M^- \cup K$ belongs to a null hypersurface L in $M^- \cup K$, the emergent null line Z^* in $M^+ \cup K$ must belong to a corresponding null hypersurface L^* in $M^+ \cup K$. We require that: $Z \cap K = Z^* \cap K$ and $L \cap K = L^* \cap K$. This serves to define Z^* uniquely in terms of Z, because $Z^* \cap K$ fixes a *point* on Z^* while $L^* \cap K$ defines a tangent 2-plane element at $Z^* \cap K$ which must be orthogonal to Z^*, thereby fixing the *direction* of Z^*.

Let us see this explicitly in terms of (4.5). A twistor Z^α, describing the null line Z according to (2.6), (2.8), satisfies

$$Z^0\,d\bar{\zeta} = Z^1\,du, \qquad Z^0\,dv = Z^1\,d\zeta \tag{4.6}$$

CHAPTER 37

$$Z^2 = i\zeta Z^1, \qquad Z^3 = i\bar{\zeta} Z^0 - iuZ^1 \tag{4.7}$$

(Here, the point P is chosen to be $Z \cap K$, so that $v = 0$.) Similarly, the starred versions of (4.6), (4.7) will also hold. In the case of (4.7), employing (4.5) we get

$$Z^{*2} = i\zeta Z^{*1}, \qquad Z^{*3} = i\bar{\zeta} Z^{*0} - i(u - r)\, Z^{*1} \tag{4.8}$$

In order to write the starred version of (4.6) in terms of du, dv, $d\zeta$ (i.e., to find the direction of Z^* given Z) we need to use the fact that Z^* is orthogonal to the same vectors lying within K, at $Z^* \cap K$, as is Z. Denoting a direction at $Z \cap K$ by $\delta u : \delta v : \delta\zeta$ in the u, v, ζ system, we have $\delta v = 0$ if the direction is to lie within K. For the direction to be orthogonal to that of Z, we have

$$\delta u\, dv + 0 . du = \delta\zeta\, d\bar{\zeta} + \delta\bar{\zeta}\, d\zeta \tag{4.9}$$

whence, by (4.6),

$$\delta u = \delta\zeta \overline{Z^0}/\overline{Z^1} + \delta\bar{\zeta} Z^0/Z^1 \tag{4.10}$$

The starred version of this yields, with (4.5),

$$\delta u - \frac{\partial r}{\partial \zeta}\delta\zeta - \frac{\partial r}{\partial \bar{\zeta}}\delta\bar{\zeta} = \delta\zeta \overline{Z^{*0}}/\overline{Z^{*1}} + \delta\bar{\zeta} Z^{*0}/Z^{*1} \tag{4.11}$$

Equations (4.10) and (4.11) must represent identical conditions on $\delta u : \delta\zeta : \delta\bar{\zeta}$ since they must give the same 2-plane element. Hence,

$$Z^0 : Z^1 = Z^{*0} + Z^{*1}\frac{\partial r}{\partial \bar{\zeta}} : Z^{*1} \tag{4.12}$$

Equations (4.7), (4.8) and (4.12) define the ratios of the $Z^{*\alpha}$ components in terms of the ratios of the Z^α components, by elimination of ζ and u. With the most convenient choice of scale factor, we can set

$$\begin{aligned} Z^{*0} &= Z^0 - Z^1 \frac{\partial r}{\partial \bar{\zeta}}, \quad & Z^{*1} &= Z^1 \\ Z^{*2} &= Z^2, & Z^{*3} &= Z^3 + iZ^1\left(r - \bar{\zeta}\frac{\partial r}{\partial \bar{\zeta}}\right) \end{aligned} \tag{4.13}$$

where $\zeta = -iZ^2/Z^1$. Setting

$$h(Z^\alpha, \bar{Z}_\alpha) \equiv |Z^1|^2 r \tag{4.14}$$

we can write (4.13) comprehensively as

$$Z^{*\alpha} = Z^\alpha + i\frac{\partial h}{\partial \bar{Z}_\alpha} \tag{4.15}$$

The transformation (4.15) has been derived for the effect due to a plane-fronted impulsive wave situated along the particular null hyperplane $v = 0$. But because of its twistor form, (4.15) will clearly apply (with suitable h) to *any* impulsive plane-fronted wave. The function $h(Z^\alpha, \bar{Z}_\alpha)$ of (4.14) is special in that it involves only the coordinates Z^2, Z^1 and their complex conjugates. Putting

$$A^\alpha = (1,0,0,0) \qquad \text{and} \qquad B^\alpha = (0,0,0,1) \tag{4.16}$$

we can express this as the fact that h is a function of

$$\bar{A}_\alpha Z^\alpha, \bar{B}_\alpha Z^\alpha; \qquad A^\alpha \bar{Z}_\alpha, B^\alpha \bar{Z}_\alpha \tag{4.17}$$

only. More generally, we could allow h to be any *real differentiable function* of variables of the type (4.17), so that h is *homogeneous of degree unity in* Z^α *and also in* $\bar{Z}_\alpha$, where we allow A^α and B^α to *any* pair of fixed null twistors satisfying $A^\alpha \bar{B}_\alpha = 0$. Note that $\partial h / \partial \bar{Z}_\alpha$ is a linear combination of A^α and B^α whence

$$\frac{\partial h}{\partial \bar{Z}_\alpha} \frac{\partial h}{\partial Z^\alpha} = 0 \tag{4.18}$$

Note, also

$$\frac{\partial^2 h}{\partial \bar{Z}_\alpha \partial Z^\alpha} = 0 \tag{4.19}$$

For the particular case considered in detail above, it is the hyperplane $v = 0$ which is generated by the null lines given by $B^\alpha + i\bar{\zeta} A^\alpha$. In the general case, these null lines would be just the generators of the null cone containing two intersecting null lines A and B [cf. (2.10)]. In the above case (4.16), A is at infinity so that the cone becomes a hyperplane. In fact, unless one of the generators of the cone *is* at infinity, (4.15) does not represent a pure impulse wave in a pseudo-Riemannian space-time but only in a *conformal* space-time. We shall ignore this distinction for the present purposes. There is, however, one special situation which is of particular note, namely when *both* A and B are at infinity, so that the null cone they define becomes the null cone at infinity itself. In this case we may regard the transformation involved in (4.5) as yielding a *supertranslation of the Bondi-Metzner-Sachs group*† and (4.15) is its twistor equivalent.

Observe that (4.18) (together with reality: $h = \bar{h}$; and homogeneity: $Z^\alpha \partial h / \partial Z^\alpha = h$) implies that

$$Z^{*\alpha} \overline{Z^*}_\alpha = Z^\alpha \bar{Z}_\alpha \tag{4.20}$$

so that the scaling we chose for $Z^{*\alpha}$, in order to arrive at (4.15), in fact preserves the twistor 'norm'. However, (4.15) does *not* preserve the twistor scalar product $Z^\alpha \bar{Y}_\alpha$. In particular, if $Z^\alpha \bar{Y}_\alpha = 0$ (so the null lines Z and Y, in M^- belong to a portion of a null cone), then in general $Z^{*\alpha} \overline{Y^*}_\alpha \neq 0$ (so that the emergent Z^*, Y^* in M^+ do not belong to a portion of a null cone). This is closely related to the non-analyticity, in Z^α, of (4.15), since it illustrates that a (shear-free) null cone picks up shear, in general, as it passes through an impulsive plane-fronted wave. (The only cases of a transformation (4.15) yielding a complex analytic twistor transformation would be given when h is bilinear in Z^α, $\bar{Z}_\alpha$. In this case the conformal curvature in the impulsive wave would vanish and M would be conformally flat.)

---

† See Sachs, R. K. (1962). *Physical Review*, **128**, 2851.

We note the important fact that whereas we originally defined the transformation $Z^\alpha \to Z^{*\alpha}$ for *null* twistors only (since the discussion was given in terms of null lines) we were led to the transformation (4.15) which applies equally to *non-null* twistors. (We may, perhaps, think of (4.19) as defining h away from N once the values of h on N are given.) Thus, the identity of a twistor can apparently be maintained even as we pass through a region of conformal curvature. The *geometrical* significance of a non-null twistor becomes altered, however. The representation of such a twistor by a Robinson congruence (cf. Section 2) cannot be maintained in the presence of conformal curvature, since this representation is based on the condition $Y^\alpha \bar{Z}_\alpha = 0$, for a null line Y to belong to the congruence $\bar{Z}$. Nevertheless the entire C-picture for M does seem to retain a significance. We may think of a single point Z in the C-picture as being assigned twistor coordinates according to two different (non-analytically related) schemes, namely that which assigns the coordinates Z^α, and that which assigns $Z^{*\alpha}$. Thus we have two *different* complex analytic structures for C and two different scalar products, depending upon whether we view it from M^- or from M^+.

The space-time M that we have just been considering is, of course, especially simple. We cannot expect to get such a complete twistor structure for a space-time M which is, perhaps, everywhere conformally curved. However, the purpose here is somewhat different. We may imagine that the effect of a region of general conformal curvature can, in some way, be *composed* of effects like that produced by an impulsive plane-fronted wave. We are thus led to consider the group $\mathscr{T}$ of transformations of twistor coordinates, which is generated by the transformations like (4.15). The invariant structure of C will then be precisely that which is left invariant by $\mathscr{T}$.

It is convenient to consider the infinitesimal transformations of the form (4.15). Denoting the infinitesimal change in the twistor coordinates Z^α by δZ^α, we can write

$$\delta Z^\alpha = i\frac{\partial H}{\partial \bar{Z}_\alpha} \tag{4.21}$$

where

$$H = H(Z^\alpha, \bar{Z}_\alpha) \tag{4.22}$$

is real and separately homogeneous of degree unity in Z^α and in $\bar{Z}_\alpha$:

$$H = \overline{H} = Z^\alpha \frac{\partial H}{\partial Z^\alpha} = \bar{Z}_\alpha \frac{\partial H}{\partial \bar{Z}_\alpha} \tag{4.23}$$

Thus, we have

$$\delta \bar{Z}_\alpha = -i \frac{\partial H}{\partial Z^\alpha} \tag{4.24}$$

It follows now that

$$\delta(Z^\alpha \bar{Z}_\alpha) = 0 \tag{4.25}$$

so that the invariance of the twistor norm follows without any other conditions on H. We need not assume that H is of the special form that was required for h, namely of being a function of variables (4.17), with its consequence (4.18). Indeed, neither (4.18) nor (4.19) can be expected to apply to a general H defining an infinitesimal element of $\mathscr{T}$. This is because the *Poisson brackets*:

$$[\psi, \chi] \equiv i \frac{\partial \psi}{\partial Z^\alpha} \frac{\partial \chi}{\partial \bar{Z}_\alpha} - i \frac{\partial \psi}{\partial \bar{Z}_\alpha} \frac{\partial \chi}{\partial Z^\alpha} \tag{4.25a}$$

do not preserve (4.18) or (4.19).

The Poisson brackets $[H,G]$ define the commutator of the two infinitesimal transformations defined by H and by G. Note that $[H,G]$ satisfies the reality and homogeneity conditions (4.23) provided H and G both satisfy these conditions. The same applies to the sum $H+G$. Thus, these operations define a Lie algebra $\mathscr{L}$. It is clear that $\mathscr{L}$ will contain the Lie algebra of infinitesimal elements of $\mathscr{T}$. (Very possibly $\mathscr{L}$ is no larger than this.) In any case it will follow that any structure on C which is invariant under $\mathscr{L}$ will also be invariant under $\mathscr{T}$, i.e., be part of the *invariant structure* of C. The converse is an open question at present.

Note that we can write (4.21), (4.24) as

$$\delta Z^\alpha = [Z^\alpha, H], \qquad \delta \bar{Z}_\alpha = [\bar{Z}_\alpha, H] \tag{4.26}$$

and, more generally, we have

$$\delta\psi = [\psi, H] \tag{4.27}$$

for any function ψ of Z^α, $\bar{Z}_\alpha$. Thus, if $[\psi, H] = 0$ for all H satisfying (4.23), then ψ is part of the invariant structure of C. By (4.25), a particular case would be $\psi = Z^\alpha \bar{Z}_\alpha$. To obtain more of the invariant structure of C, we must consider 'tensor fields' on C. In particular, there are certain invariant *differential forms*† on C. Noting that, by (4.21),

$$\delta(dZ^\alpha) = d\left(i\frac{\partial H}{\partial \bar{Z}_\alpha}\right) = i\frac{\partial^2 H}{\partial Z^\beta\,\partial \bar{Z}_\alpha}\,dZ^\beta + i\frac{\partial^2 H}{\partial \bar{Z}_\beta\,\partial \bar{Z}_\alpha}\,d\bar{Z}_\beta \tag{4.28}$$

and using the fact that $\partial H/\partial Z^\beta$ and $\partial H/\partial \bar{Z}_\beta$ are respectively homogeneous of degrees one and zero in $\bar{Z}_\alpha$, we obtain

$$\delta(\bar{Z}_\alpha\,dZ^\alpha) = 0 \tag{4.29}$$

Thus $\bar{Z}_\alpha dZ^\alpha$ is part of the invariant structure of C and, hence, so also is its exterior derivative $d(\bar{Z}_\alpha dZ^\alpha) = d\bar{Z}_\alpha \wedge dZ^\alpha = 0$, i.e.

$$\delta(d\bar{Z}_\alpha \wedge dZ^\alpha) = 0 \tag{4.30}$$

The quantity $d\bar{Z}_\alpha \wedge dZ^\alpha$ is the usual invariant surface element associated with the 'Hamiltonian' equations (4.21), (4.24). By taking exterior products of these forms, higher degree invariant quantities can be (rather trivially) generated, e.g. $Z^\alpha d\bar{Z}_\alpha \wedge dZ^\beta \wedge d\bar{Z}_\beta$; $\bar{Z}_\alpha Z^{(\alpha} dZ^{\beta)} \wedge d\bar{Z}_\beta$. The 7-form $Z^\alpha d\bar{Z}_\alpha \wedge dZ^\beta \wedge d\bar{Z}_\beta \wedge dZ^\gamma \wedge d\bar{Z}_\gamma \wedge dZ^\mu \wedge d\bar{Z}_\mu$ that can be built in this way is dual to $Z^\alpha \partial/\partial Z^\alpha$. The invariance of this operator follows more directly from

$$\left[Z^\alpha \frac{\partial \psi}{\partial Z^\alpha}, H\right] = Z^\alpha \frac{\partial}{\partial Z^\alpha}[\psi, H] \tag{4.31}$$

which we can interpret as

$$\delta\left(Z^\alpha \frac{\partial}{\partial Z^\alpha}\right) \equiv \left(Z^\alpha \frac{\partial}{\partial Z^\alpha}\right)\delta \tag{4.32}$$

This is the condition for invariance of the operator $Z^\alpha \partial/\partial Z^\alpha$ under δ.

The most important quantities belonging to the invariant structure of C can be collected together as follows:

$$\bar{Z}^\alpha Z_\alpha \tag{4.33}$$

$$Z^\alpha d\bar{Z}_\alpha, \qquad \bar{Z}_\alpha dZ^\alpha \tag{4.34}$$

$$dZ^\alpha \wedge d\bar{Z}_\alpha \tag{4.35}$$

$$Z^\alpha \frac{\partial}{\partial Z^\alpha}, \qquad \bar{Z}_\alpha \frac{\partial}{\partial \bar{Z}_\alpha} \tag{4.36}$$

It is of interest that all these quantities have *some* direct significance in the M-picture. It is the vanishing of (4.33) which states that the twistor Z^α represents a *real null line* Z in M. The division of C into the three parts C^-, N, C^+ is invariant, being given by $Z^\alpha \bar{Z}_\alpha > 0, = 0, < 0$, respectively. However, the points of C^+ and C^- do not appear to have any very direct invariant geometrical interpretation in M. As for (4.34), we may think of Z^α and $Z^\alpha + dZ^\alpha$ as defining two infinitesimally neighbouring points Z and Z' in C. If Z and Z' are both on N, then in the M-picture, Z and Z' are infinitesimally neighbouring null lines and we have $Z^\alpha \bar{Z}_\alpha = 0$ and $Z^\alpha d\bar{Z}_\alpha + dZ^\alpha \bar{Z}_\alpha = 0$ (neglecting quantities of second order in dZ^α). Thus, $Z^\alpha d\bar{Z}_\alpha$ is pure imaginary. Now, the imaginary part of $Z^\alpha d\bar{Z}_\alpha$ is *positive* or *negative* according as the null line Z' lies just to the *future* or just to the *past* of Z. If $Z^\alpha d\bar{Z}_\alpha = 0$, then a *null hypersurface* can contain both Z and Z' as neighbouring generators. Now consider (4.35). We may think of Z^α, $Z^\alpha + dZ^\alpha$ and $Z^\alpha + d'Z^\alpha$ as defining three neighbouring null lines Z, Z' and Z'' in M provided these three twistors are all null. Let us suppose that this is so, and also that any pair of Z, Z', Z'' can belong to a null hypersurface, i.e. $Z^\alpha \bar{Z}_\alpha = 0$, $Z^\alpha d\bar{Z}_\alpha = 0$, $Z^\alpha d'\bar{Z}_\alpha = 0$. Then the condition that all three of Z, Z', Z'' can belong to one null hypersurface (i.e. that there be no net rotation of Z, Z', Z'' about one another) is $dZ^\alpha d'\bar{Z}_\alpha - d'Z^\alpha d\bar{Z}_\alpha = 0$, which in the notation of differential forms reads $dZ^\alpha \wedge d\bar{Z}_\alpha = 0$. The form $dZ^\alpha \wedge d\bar{Z}_\alpha$ defines a *symplectic* structure on C. Finally, the invariance of (4.36) is a necessary prerequisite for the C-picture to make geometrical sense at all. For, we have tacitly assumed that it is legitimate to think of C as a six-real-dimensional manifold, the points of which are defined by the *ratios* of the complex coordinates Z^α. The fact that the relation between Z^α and λZ^α is preserved under any transformation in $\mathscr{L}$ is implicit in the fact that the (degree of) homogeneity in Z^α of a function $\psi(Z^\alpha, \bar{Z}_\alpha)$ is preserved under $\mathscr{L}$. The functions homogeneous in Z^α are simply the eigenfunctions of the invariant operator $Z^\alpha \partial/\partial Z^\alpha$.

---

† See footnote on page 92.

The above discussion has been centered on considerations of infinitesimal transformations. We may also consider finite transformations $Z^\alpha \to Z^{*\alpha}$ belonging to $\mathscr{T}$. For such transformations we must expect that the expression of the Poisson brackets in terms of Z^α, $\bar{Z}_\alpha$ will hold also for $Z^{*\alpha}$, $\overline{Z^*}_\alpha$. We have

$$[Z^\alpha, Z^\beta] = 0 = [\bar{Z}_\alpha, \bar{Z}_\beta], \qquad [Z^\alpha, \bar{Z}_\beta] = i\delta_\beta{}^\alpha \tag{4.37}$$

Thus also

$$[Z^{*\alpha}, Z^{*\beta}] = 0 = [\overline{Z^*}_\alpha, \overline{Z^*}_\beta], \qquad [Z^{*\alpha}, \overline{Z^*}_\beta] = i\delta_\beta{}^\alpha \tag{4.38}$$

These equations can be stated

$$\frac{\partial Z^{*\alpha}}{\partial Z^\beta} = \frac{\partial \bar{Z}_\beta}{\partial \overline{Z^*}_\alpha}, \qquad \frac{\partial Z^{*\alpha}}{\partial \bar{Z}_\beta} = -\frac{\partial Z^\beta}{\partial \bar{Z}_\alpha} \tag{4.39}$$

or, equivalently

$$dZ^{*\alpha} \wedge d\overline{Z^*}_\alpha = dZ^\alpha \wedge d\bar{Z}_\alpha \tag{4.40}$$

If, in addition, we assume that $Z^{*\alpha}$ is homogeneous of degree one in Z^α and of degree zero in $\bar{Z}_\alpha$, i.e.

$$Z^{*\alpha}\frac{\partial}{\partial Z^{*\alpha}} \equiv Z^\alpha \frac{\partial}{\partial Z^\alpha} \tag{4.41}$$

then it follows that

$$Z^{*\alpha}\,\overline{Z^*}_\alpha = Z^\alpha \bar{Z}_\alpha, \qquad Z^{*\alpha}\,d\overline{Z^*}_\alpha = Z^\alpha\, d\bar{Z}_\alpha \tag{4.42}$$

also.

To sum up, the invariant structure obtained here for the C-picture is defined by a symplectic structure given by $dZ^\alpha \wedge d\bar{Z}_\alpha$, where the 'homogeneity' operator $Z^\alpha \partial/\partial Z^\alpha$ is also invariant. Equivalently, $Z^\alpha \bar{Z}_\alpha$ and $Z^\alpha d\bar{Z}_\alpha$ are invariant. If we restrict our attention to null twistors, i.e. to N, this structure describes a (conformally invariant) geometry of null geodesics in M, which heeds only the null geodesics as a whole and does not refer to points on these null geodesics. The non-null twistors appear to play a 'catalytic' role in simplifying the description of the geometry of N. The structure so obtained for N (and for C) is of a 'universal' nature; that is to say, it does not reflect the local metric (or conformal) structure of M in any way. (We can see this by removing a portion of M and joining it smoothly on to another space-time manifold, e.g. to flat space-time. The invariant structure of C does not change.) To represent local structure of M—and, in particular, its points—we would have to refer to some additional (generally non-local) structure for N over and above its invariant structure.

5. *A Hilbert Space*

The nature of the twistor transformations induced by the presence of (conformal) curvature in the M-picture strongly suggests that, in the passage to a quantum theory, we should identify a quantum operator $\bar{Z}_\alpha$ with some multiple of an operator $\partial/\partial Z^\alpha$. In order to realise such an identification, we must obtain the vectors of the appropriate Hilbert space on which these operators are to act. It is a remarkable fact that one such space is *already* at hand, having been previously required for the twistor description of zero rest-mass fields given in Section 3. This is the space of the analytic functions $f(Z^\alpha)$ which are employed in the contour integrals (3.10). In order to get a Hilbert space, however, we shall require a definition of norm (or of scalar product) for these functions. I shall not enter into all the details involved with this here, some of which involve topological questions concerning the singularity sets, but merely give the *formal* expression for the Hilbert space scalar product. These detailed matters will be discussed elsewhere.

Let $f(Z^\alpha)$, $g(Z^\alpha)$ be two functions of the complex variables Z^0, Z^1, Z^2, Z^3 which are homogeneous of degree $-2s-2$ and analytic in suitable domains. For example, we could choose domains of analyticity of the type considered towards the end of Section 3, where the regions of $N \cup C^+$ which were excluded, were two disjoint closed sets (each of which intersected every line in $N \cup C^+$). The scalar product is then

$$\langle g|f\rangle = ik\Gamma(2-2s)\oint \bar{g}(W_\alpha)f(Z^\alpha)(W_\beta Z^\beta)^{2s-2}\mathscr{W}\wedge\mathscr{Z}. \qquad (5.1)$$

where k is a real numerical constant, $\mathscr{W}$ and $\mathscr{Z}$ are differentials given by

$$\mathscr{W} = \tfrac{1}{6}\epsilon^{\alpha\beta\gamma\delta}W_\alpha dW_\beta \wedge dW_\gamma \wedge dW_\delta \qquad (5.2)$$

$$\mathscr{Z} = \tfrac{1}{6}\epsilon_{\alpha\beta\gamma\delta}Z^\alpha dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \qquad (5.3)$$

(the ϵ's being the usual Levi-Civita symbols), and where the (six-dimensional) region of integration (contour) is compact, and surrounds the singularities of f and g, and the region $W_\beta Z^\beta = 0$ in a suitable way. When $s = 1, 1\frac{1}{2}, 2, \ldots$ the expression (5.1) is not defined as it stands, the value of the integral being zero and the multiplying factor infinite. In these cases we can, however, assign a meaning to (5.1) as the result of a *limit* process applied to s, where s is taken as a continuous variable. This will be discussed a little more shortly.

We have to establish that (5.1) is invariant under homologous deformations of the contour over regions of analyticity of the integrand. Thus† we must verify that the exterior derivative of the expression after the integral sign vanishes. As a lemma towards achieving this, consider any function $p(Z^\alpha; u_1, \ldots, u_h)$ which is analytic and homogeneous of degree -4 in the Z^α and which depends differentiably on the parameters $u_1, \ldots, u_h$. Then

$$d(p\mathscr{Z}) = \frac{\partial p}{\partial u_i}du_i \wedge \mathscr{Z} \qquad (5.4)$$

---

† For an account of exterior calculus, see, for example, Hodge, W. V. D. (1952). *Theory and Applications of Harmonic Integrals.* Cambridge University Press; Flanders, H. (1963). *Differential Forms, with Applications to the Physical Sciences.* Academic Press, New York and London.

To demonstrate the validity of (5.4), we must show that the terms involving derivatives of the Z^α cancel out. We have

$$d(p\mathscr{X}) = p\,d\mathscr{X} + \frac{\partial p}{\partial Z^\mu} dZ^\mu \wedge \mathscr{X} + \frac{\partial p}{\partial u_i} du_i \wedge \mathscr{X} \tag{5.5}$$

and

$$\begin{aligned} \frac{\partial p}{\partial Z^\mu} dZ^\mu \wedge \mathscr{X} &= \frac{1}{6}\frac{\partial p}{\partial Z^\mu} Z^\alpha \epsilon_{\alpha\beta\gamma\delta}\, dZ^\mu \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \\ &= \frac{1}{4!6}\frac{\partial p}{\partial Z^\mu} Z^\alpha \epsilon_{\alpha\beta\gamma\delta}\, \epsilon^{\mu\beta\gamma\delta}\, \epsilon_{\lambda\rho\sigma\tau}\, dZ^\lambda \wedge dZ^\rho \wedge dZ^\sigma \wedge dZ^\tau \\ &= \frac{1}{4} Z^\mu \frac{\partial p}{\partial Z^\mu} \cdot \frac{1}{6} \epsilon_{\lambda\rho\sigma\tau}\, dZ^\lambda \wedge dZ^\rho \wedge dZ^\sigma \wedge dZ^\tau \end{aligned} \tag{5.6}$$

by the skew-symmetry of $dZ^\mu \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta$ and the properties of the ϵ's. Now, by the homogeneity of p we have

$$\tfrac{1}{4} Z^\mu \frac{\partial p}{\partial Z^\mu} = -p \tag{5.7}$$

Furthermore,

$$d\mathscr{X} = \tfrac{1}{6}\epsilon_{\lambda\rho\sigma\tau}\, dZ^\lambda \wedge dZ^\rho \wedge dZ^\sigma \wedge dZ^\tau \tag{5.8}$$

Combining (5.5), (5.6), (5.7) and (5.8), we obtain (5.4) as required.

The expression $q = \bar{g}(W_\alpha) f(Z^\alpha)(W_\beta Z^\beta)^{2s-2}$ which occurs in (5.1) is analytic and homogeneous of degree -4 in Z^α and also in W_α. If we apply (5.4), with the coefficients of the 4-form $q\mathscr{W}$ in place of p, we see that $d(q\mathscr{W} \wedge \mathscr{X})$ contains no term of degree four in the dZ's. Similarly, $d(q\mathscr{W} \wedge \mathscr{X})$ contains no term of degree four in the dW's. Hence $d(q\mathscr{W} \wedge \mathscr{X}) = 0$, as required. This shows that the integral (5.1) does not depend on the exact position of the contour, but only on its homology class in the region of analyticity of q. It is the singularity regions of $f(Z^\alpha)$, $\bar{g}(W_\alpha)$ and $(Z^\alpha W_\alpha)^{-2+2s}$ which must prevent the contour from being homologous to zero. Otherwise the integral in (5.1) would vanish. Something of this nature is, of course, to be expected, since it is the *separation of singularities* of f, in (3.10), which gives rise to a zero rest-mass field.

It may be verified directly that the expression (5.1) arises from precisely the usual definition of scalar product for zero rest-mass fields, in the cases $s = 0$, $s = \frac{1}{2}$. (The argument will be given elsewhere.) The cases $s = 1$, $1\frac{1}{2}$, 2, ... are less straightforward. Here the factor $(Z^\rho W_\rho)^{-2+2s}$ is no longer singular at $Z^\rho W_\rho = 0$, so the contour can now be deformed across this region. In fact, for the type of field considered here, the contour can always be deformed to a point, so that the integral in (5.1) vanishes. This is compensated by the pole in

$\Gamma(2-2s)$. To obtain the meaning of (5.1) in these circumstances, we may resort to a limiting argument. If f and g are mutiplied by suitable factors homogeneous of degree ϵ, and $\Gamma(2-2s)(Z^\mu W_\mu)^{-2+2s}$ is replaced by $\Gamma(2-2s+\epsilon)(Z^\mu W_\mu)^{-2+2s-\epsilon}$, then (5.1) becomes well-defined. Allowing ϵ to tend to zero and the multiplying factors to tend to unity, we can obtain a finite meaning for (5.1) when $s = 1, 1\frac{1}{2}, 2, \ldots$ also.

Let us see this explicitly, for a particular choice of multiplying factor. Choose A_α, B^β and the contour suitably, so that the contour avoids the region $A_\alpha Z^\alpha B^\beta W_\beta = 0$. (This appears to be possible under 'normal' circumstances. For more exotic functions f and g it might be necessary to choose more complicated multiplying factors.) Then

$$\begin{aligned}\langle g|f\rangle &= \lim_{\epsilon\to 0}\{ik\Gamma(2-2s+\epsilon)\times \\ &\quad\times \oint \bar{g}(W_\alpha)(B^\beta W_\beta)^\epsilon f(Z^\alpha)(A_\gamma Z^\gamma)^\epsilon (Z^\mu W_\mu)^{2-2s-\epsilon}\,\mathscr{W}\wedge\mathscr{Z}\} \\ &= \frac{ik(-1)^{2s-2}}{(2s-2)!}\times \qquad\qquad (5.9) \\ &\quad\times \oint \bar{g}(W_\alpha) f(Z^\alpha)(Z^\mu W_\mu)^{2s-2}\ln\left\{\frac{B^\beta W_\beta A_\gamma Z^\gamma}{Z^\rho W_\rho}\right\}\mathscr{W}\wedge\mathscr{Z}\end{aligned}$$

by l'Hospital's rule, provided $s = 1, 1\frac{1}{2}, 2, \ldots$. The fact that (5.9) does not change as A_α or B^α are varied is again a consequence of the vanishing of the integral in (5.1) (but with a slightly modified function in place of f or g), as is readily verified. With this interpretation [e.g. (5.9)] for the scalar product, it can be shown that also in the cases $s = 1, 1\frac{1}{2}, 2, \ldots$ the expression (5.1) agrees with the usual definition in terms of fields. (The argument will be given elsewhere.) The more complicated behaviour for $s = 1, 1\frac{1}{2}, 2, \ldots$, than for $s = 0, \frac{1}{2}$ seems to be related to the fact that a similar increase in complication occurs in the usual formalism at this point owing to the fact that the number operator becomes non-local (or involves potentials) when $s > \frac{1}{2}$. It is also related to the existence of conserved integrals of the field (charge for $s = 1$; mass, momentum and angular momentum for $s = 2$) as we shall see shortly. Note, particularly, that the scalar product (5.1) is *conformally invariant*† (and *a fortiori* Poincaré invariant) because of twistor covariance. The definition can also be applied to *non-half-integral* spin s. The positive definiteness of (5.1) (for positive frequency fields) will be discussed elsewhere, as will the *details* of the definition of the Hilbert space (e.g. two different functions f may correspond to the same Hilbert space vector).

---

† Compare Gross, L. (1963). *Journal of Mathematics and Physics*, **5**, 687.

The object, now, is to show that the operators Z^α and $\overleftarrow{\partial}/\partial\bar{Z}_\alpha$ are equivalent to each other in their effect on the Hilbert space. That is to say, for any two functions $f(Z^\alpha)$, $g(Z^\alpha)$ of the type we have been considering, but with f and g homogeneous of respective degrees $-2s-3$ and $-2s-2$, we have (A_α being constant),

$$A_\alpha\langle g|Z^\alpha|f\rangle = \langle g|A_\alpha Z^\alpha f\rangle = \langle \bar{A}^\alpha \frac{\partial g}{\partial Z^\alpha}\Big|f\rangle = A_\alpha \langle g\Big|\frac{\overleftarrow{\partial}}{\partial \bar{Z}_\alpha}\Big|f\rangle \tag{5.10}$$

We must verify the middle equality in (5.10). Thus, we have to show that, for all s,

$$\oint A_\lambda(W_\beta Z^\beta)^{2s-2}f(Z^\alpha)\left\{Z^\lambda(2s-1)\bar{g}(W_\alpha) + W_\rho Z^\rho \frac{\partial\bar{g}(W_\alpha)}{\partial W_\lambda}\right\}\mathscr{W}\wedge\mathscr{Z} = 0 \tag{5.11}$$

[since $(2s-1)\Gamma(1-2s) = -\Gamma(2-2s)$]. (If $s=\frac{1}{2}, 1, 1\frac{1}{2}, \ldots$, then (5.11) 'automatically' vanishes and does not, in itself, imply (5.10) for these s values. But it will follow that (5.10) holds if we can verify (5.11) for *all* real s, since then a limiting argument on s will apply.) Now‡ (5.11) will hold if we can show that the expression under the integral sign is of the form $d\mathscr{X}$ for some 5-form $\mathscr{X}$. Set

$$\mathscr{X} = -\tfrac{1}{2}A_\lambda f(Z^\gamma)\,\bar{g}(W_\alpha)\,(W_\beta Z^\beta)^{2s-1}\,\epsilon^{\lambda\mu\nu\sigma}\,W_\mu\, dW_\nu\wedge dW_\sigma\wedge\mathscr{Z}$$

We wish to calculate $d\mathscr{X}$. By (5.4), we need only consider the terms involving derivatives with respect to W's. The calculation is straightforward if the identity

$$\begin{aligned}\epsilon^{\lambda\mu\nu\sigma}\,dW_\rho\wedge dW_\nu\wedge dW_\sigma &= \tfrac{1}{6}\epsilon^{\lambda\mu\nu\sigma}\,\epsilon_{\tau\rho\nu\sigma}\,\epsilon^{\tau\psi\chi\phi}\,dW_\psi\wedge dW_\chi\wedge dW_\phi \\ &= \tfrac{1}{3}(\delta_\rho{}^\mu\,\epsilon^{\lambda\psi\chi\phi} - \delta_\rho{}^\lambda\,\epsilon^{\mu\psi\chi\phi})\times \\ &\quad\times dW_\psi\wedge dW_\chi\wedge dW_\phi\end{aligned} \tag{5.13}$$

is used, the terms in $A_\lambda\epsilon^{\lambda\mu\nu\sigma}dW_\mu\wedge dW_\nu\wedge dW_\sigma$ cancelling out. The result is that $d\mathscr{X}$ turns out to be precisely the term under the integral sign in (5.11), as required.

On the basis of (5.10) we can now write

$$Z^\alpha = \frac{\overleftarrow{\partial}}{\partial\bar{Z}_\alpha} \tag{5.14}$$

where these are to be read as operators in the Hilbert space. The conjugate of (5.14) gives us

$$\bar{Z}_\alpha = \frac{\overrightarrow{\partial}}{\partial Z^\alpha} \tag{5.15}$$

---

‡ See footnote on p. 92

(Planck's constant has been implicitly absorbed into the Z^α by the definition (5.1), since k has been chosen to be independent of s. This, then, implies a particular scaling for the $\phi_{A\ldots L}$ fields if the formulae of Section 3 are strictly adhered to.) We now have the commutation relation

$$\bar{Z}_\alpha Z^\beta - Z^\beta \bar{Z}_\alpha = \delta_\alpha{}^\beta \tag{5.16}$$

which may be compared with (4.37) (we, of course, also have $Z^\alpha Z^\beta = Z^\beta Z^\alpha$, $\bar{Z}^\alpha \bar{Z}^\beta = \bar{Z}^\beta \bar{Z}^\alpha$). This suggests a classical-quantum correspondence (Dirac, 1958)

$$i[\psi, \chi] = \psi\chi - \chi\psi \tag{5.17}$$

when ψ and χ are to be suitable quantum operators.

Note that the 'homogeneity operator' $Z^\alpha \partial/\partial Z^\alpha$ is just $Z^\alpha \bar{Z}_\alpha$. Since the eigenvalues of this operator are the values $-2s - 2$, by (5.16) we can write

$$-\tfrac{1}{4}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha) = s \tag{5.18}$$

for the spin operator. The trace-free 'Hermitian' operator

$$\begin{aligned} E_\beta{}^\alpha &= Z^\alpha \bar{Z}_\beta - \tfrac{1}{4}\delta_\beta{}^\alpha Z^\gamma \bar{Z}_\gamma \\ &= \bar{Z}_\beta Z^\alpha - \tfrac{1}{4}\delta_\beta{}^\alpha \bar{Z}_\gamma Z^\gamma \end{aligned} \tag{5.19}$$

generates the *infinitesimal conformal transformations* of M (since if $p_\alpha{}^\beta$ is 'Hermitian' and trace-free: $\bar{p}_\beta{}^\alpha = p_\beta{}^\alpha$, $p_\alpha{}^\alpha = 0$; then the infinitesimal twistor transformation $Z^\alpha \to Z^\beta(\delta_\beta{}^\alpha + i\epsilon p_\beta{}^\alpha)$ is defined by the operator $i\epsilon p_\beta{}^\alpha E_\alpha{}^\beta = i\epsilon Z^\beta p_\beta{}^\alpha \partial/\partial Z^\alpha$, neglecting ϵ^2). Thus, the fifteen components of $E_\beta{}^\alpha$ include the *energy*, the three components of *momentum* and the six components of relativistic *angular momentum*, in addition to the five extra conservation laws which arise from the conformal invariance. If we wish to single out only the ten components of energy-momentum and angular momentum, then we employ the operator

$$F^{\alpha\beta} = E_\gamma^{(\alpha} I^{\beta)\gamma} \tag{5.20}$$

where $I^{\alpha\beta}$ is the (fixed) metric twistor (Penrose, 1967a) (whose only non-zero components, in the coordinate system of Section 2, are $I^{23} = -I^{32} = 1$; also, for $I_{\alpha\beta}$, just $I_{01} = -I_{10} = 1$ are non-zero).

It is of interest to see what the M-picture interpretation of the operators Z^α and $\partial/\partial Z^\alpha$ amounts to. Let $\phi_{AB\ldots L}$ be the spin s field corresponding to $f(Z^\alpha)$ as in Section 3. Then the spin $s - \frac{1}{2}$ field corresponding to $T_\beta Z^\beta f(Z^\alpha)$, where T_β is some constant twistor, is just

$$\psi_{B\ldots L} = \phi_{AB\ldots L}(-i\kappa^A + x^{AB'}\rho_{B'}) = -i\phi_{AB\ldots L}\tau^A \tag{5.21}$$

where, as in (2.14), the constant spinors κ^A, $\rho_{B'}$ are defined by $T_0 = -\kappa^1$, $T_1 = \kappa^0$, $T_2 = \rho_{1'}$, $T_3 = -\rho_{0'}$. Similarly, the spin $s + \frac{1}{2}$ field corresponding to $\bar{T}^\beta \partial f / \partial Z^\beta$ is

$$\chi_{AB\ldots LM} = i(2s+1)\bar{\rho}_{(M}\phi_{AB\ldots L)} - (\bar{\kappa}^{Q'} - ix^{PQ'}\bar{\rho}_P)\nabla_{Q'M}\phi_{AB\ldots L} \quad (5.22)$$
$$= -(s+\tfrac{1}{2})\{\nabla_{Q'(M}\tau^{Q'}\}\phi_{AB\ldots L)} - \bar{\tau}^{Q'}\nabla_{Q'M}\phi_{AB\ldots L}$$

One may verify directly that the $\psi_{B\ldots L}$ and $\chi_{A\ldots M}$ of (5.21), (5.22) satisfy the zero rest-mass equation (3.3), by virtue of (2.15): $\nabla_{P'(A}\tau_{B)} = 0$. In (5.22), $\phi_{A\ldots L}$ acts as a kind of potential field for $\chi_{A\ldots M}$. Particularly in the case $\bar{\rho}_P = 0$ (so τ^A is constant), this type of potential has been studied earlier (Penrose, 1965).

Finally, we may try to interpret, in the present formalism, the (conserved) 2-surface integrals of electromagnetic field ($s = 1$) or of linearised gravitational field ($s = 2$) which respectively define the total charge or total *mass*, *momentum*, and *angular momentum* of a *source* for the field. For this purpose we must allow ourselves to consider singularity regions for the function f which are more extensive than arose for the free-wave fields we had been considering previously. For the case of charge we have

$$Q = k_1 \oint f(Z^\alpha)\,\mathscr{Z} \quad (5.23)$$

where f is homogeneous of degree -4; and for mass, momentum and angular momentun,

$$G^{\alpha\beta} = k_2 \oint Z^\alpha Z^\beta f(Z^\alpha)\,\mathscr{Z} \quad (5.24)$$

where f is homogeneous of degree -6. Here k_1 and k_2 are constants, and the three-dimensional closed contour is chosen suitably so as to correspond to a region of free field surrounding the sources. Expressions (5.23) and (5.24) can both be obtained by directly translating the usual expressions in terms of the fields. (The arguments will be given elsewhere.) For free waves, the integrals (5.23), (5.24), of course, both must *vanish*. From this, we can see at once that the integral in (5.1) must necessarily vanish for such waves ($s = 1, 2$)—hence the necessity of the limiting procedure to define scalar product ($s = 1, 1\frac{1}{2}, 2, \ldots$).

Note that (5.24) shows how the mass, momentum and angular momentum integrals can all be reduced to 'charge' integrals for certain spin-1 fields constructible from the given spin-2 field ϕ_{ABCD}. For, if we multiply (5.24) by a constant symmetric twistor $S_{\alpha\beta}$ then (5.24) reduces to (5.23), with $f(Z^\alpha)$ replaced by $S_{\alpha\beta}Z^\alpha Z^\beta f(Z^\alpha)$. In the

M-picture, this corresponds to calculating the 'charge' integral for the spin-1 field given by

$$\theta_{AB} = -\frac{k_2}{k_1}\phi_{ABCD}\,\sigma^{CD} \tag{5.25}$$

where σ^{CD} is the symmetric spinor, satisfying $\nabla^{P'(Q}\sigma^{CD)} = 0$, which corresponds to $S_{\alpha\beta}$ as in (2.17). For sources for which ϕ_{ABCD} can be derived from a potential, the 'magnetic' parts of (5.24) vanish. If $S_{\alpha\gamma} I^{\beta\gamma}$ is 'Hermitian' (i.e. $= \overline{S}^{\beta\gamma} I_{\alpha\gamma}$), then this means that there is zero 'magnetic charge' for θ_{AB} (as would follow if θ_{AB} is derivable from a potential). This condition, in terms of (5.24), is the 'Hermiticity' relation

$$G^{\alpha\gamma} I_{\beta\gamma} = \bar{G}_{\beta\gamma} I^{\alpha\gamma} \tag{5.26}$$

Equation (4.26) is the condition for $G^{\alpha\beta}$ $(= G^{\beta\alpha})$ to be of the form

$$G^{\alpha\beta} = B^{(\alpha}_{\gamma} I^{\beta)\gamma} \tag{5.27}$$

where $B_{\gamma}{}^{\alpha} = \bar{B}_{\gamma}{}^{\alpha}$, $B_{\alpha}{}^{\alpha} = 0$.

We may compare (5.20) with (5.27). The $F^{\alpha\beta}$ of (5.20) describes the total mass, etc., in the sense of *inertial* mass or *energy* content of the field. However, $G^{\alpha\beta}$ describes the total *active* mass, etc., as it appears as the *source* for the particular spin-2 field ϕ_{ABCD} [defined by the f of (5.24)]. This suggests a possible way that the actual *field equations* of general relativity might eventually be incorporated into the present formalism. For ϕ_{ABCD} to describe the gravitational field, we should expect something like

$$\langle F^{\alpha\beta} \rangle = G^{\alpha\beta} \tag{5.28}$$

A complication which will naturally arise, springs from the fact that ϕ_{ABCD} must, itself, contribute to $\langle F^{\alpha\beta} \rangle$. This would result in non-linearities of a type familiar in general relativity theory.

6. *Conclusions*

The twistor formalism appears to afford considerable scope for the expression of basic physical processes, several aspects of physics fitting unexpectedly naturally into the twistor framework. The development given here is an approximately 'historical' one. It was the desire to make the formalism fit in with general relativity which suggested the identification $\bar{Z}_{\alpha} = \partial/\partial Z^{\alpha}$ as a basis for an approach to quantization. This in turn led to the correct twistor expression for the zero rest-mass field Hilbert space scalar product (which had hitherto

CHAPTER 37

proved to be elusive). The identification emerged as being *consistent* with this scalar product. It is thus very tempting to believe that a link between space-time curvature and quantum processes may be supplied by the use of twistors. Then, roughly speaking, it is the continual slight 'shifting' of the interpretations of the quantum (twistor) operators which results in the curvature of space-time.

References

Brauer, R. and Weyl, H. (1935). *American Journal of Mathematics*, **57**, 425.

Brinkmann, W. H. (1923). *Proceedings of the National Academy of Sciences of the United States of America*, **9**, 1. For further references and details, see Ehlers, J. and Kundt, W. (1962). In: *Gravitation: an Introduction to Current Research*, ed. Witten, L. John Wiley and Sons, New York and London.

Brinkmann, W. H. (1925). *Mathematische Annalen*, **94**, 119.

Bruhat, Y. (1959). *Comptes Rendus de l'Académie des Sciences*, **248**, 1782.

Cartan, E. (1914). *Annales de l'École Normale Supérieùre*, **31**, 263.

Dirac, P. A. M. (1936). *Proceedings of the Royal Society*, **A155**, 447.

Dirac, P. A. M. (1958). *The Principles of Quantum Mechanics*, 4th ed. Clarendon Press, Oxford.

Fierz, M. (1940). *Helvetica physica acta*, **13**, 45.

Grgin, E. (1966). Ph.D. Thesis, Syracuse University.

Hepner, W. A. (1962). *Nuovo cimento*, **26**, 351.

Kastrupp, H. A. (1962). *Physics Letters (Amsterdam)*, **3**, 78.

Kastrupp. H. A. (1966). *Physical Review*, **142**, 1060.

Kerr, R. P. (unpublished); see Penrose (1967a).

Murai, Y. (1953). *Progress of Theoretical Physics (Kyoto)*, **9**, 147.

Murai, Y. (1954). *Progress of Theoretical Physics (Kyoto)*, **11**, 441.

Murai, Y. (1958). *Nuclear Physics*, **6**, 489.

Penrose, R. (1963). Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity. In: *A.R.L. Tech. Documentary Rep.* 63–56, ed. Bergmann, P. G.

Penrose, R. (1965). *Proceedings of the Royal Society*, **A284**, 159.

Penrose, R. (1967a). *Journal of Mathematics and Physics*, **8**, 345.

Penrose, R. (1967b). *An Analysis of the Structure of Space-Time.* (Adams prize essay; Princeton.)

Penrose, R. (1968). (To be published).

Robinson, I. (1958). *Report to the Eddington Group.* Cambridge.

Segal, I. E. (1967). *Proceedings of the National Academy of Sciences of the United States of America*, **57**, 194.

van der Waerden, B. L. (1929). *Nachrichten von der Gesellschaft der Wissenschaften Göttingen*, **100**, 1.

38

ON GRAVITATIONAL COLLAPSE

Contemporary Physics, Vol. 1 Trieste Symposium, 1968 (International Atomic Energy Agency, Vienna 1969) pp. 545–56

218 On gravitational collapse. *Contemporary Physics, Vol. 1 Trieste Symposium,* 1968 (International Atomic Energy Agency, Vienna 1969) pp. 545–56

ON GRAVITATIONAL COLLAPSE

R. PENROSE
Department of Mathematics,
Birkbeck College,
London, United Kingdom

Abstract

ON GRAVITATIONAL COLLAPSE. The gravitational collapse of a spherically symmetrical body, to within its Schwarzschild radius, is discussed in qualitative terms. The concept of a "trapped surface" is presented, in order to deal with asymmetrical cases. The effect of rotation is illustrated by reference to the Kerr solution. Two cases, namely of "small" and "large" angular momentum, are contrasted. In the latter case no trapped surface (and no event horizon) arises.

A new theorem, due jointly to S.W. Hawking and the present author, is described, which effectively incorporates and generalizes the previous theorems predicting space-time singularities. In particular, if a physically acceptable space-time contains a trapped surface (or, alternatively, if it is spatially closed), then on the basis of Einstein's equations (with $\lambda = 0$), it must possess a singularity.

I have taken it that I am to present a "point of view" on relativity. In fact, I think I shall present not one but *two* (more or less orthogonal) points of view - neither of which really adequately represents my own viewpoint.

The first point of view is that of the *pure relativist*. The pure relativist treats general relativity as a branch of differential geometry. He is really more a pure mathematician than a physicist and is not normally concerned with the relation of his theory to observation or experiment. He strives for elegance in his arguments and regards this as an end in itself. He has a strong distaste for co-ordinates in general, although on occasion he reluctantly finds he has the need to use them. He is an extreme sceptic as regards physical concepts, even those which have historically been regarded as lying at the root of his subject. He cannot, for example, comprehend the meaning of such terms as "general covariance", "the strong equivalence principle" (although the "weak equivalence principle" he can probably manage) or "Mach's principle". He cannot, of course, understand the concept of "force" or of "potential energy" but a "*total* energy-momentum" concept he can, with great effort, just about accept.

The second point of view I wish to present can be referred to as that of the *gravitational optimist* (although some people might feel that "pessimist" might be a more appropriate term). The gravitational optimist believes that general relativity has *fundamental* importance to the physics of our world. He believes, in particular, that the observational aspects of the theory will not always be restricted to a few very minor effects like the perihelion advance of Mercury; that cosmological interest in the theory will not ultimately be restricted to a few simple smoothed-out symmetrical models. Moreover, he would like to believe that the effects of space-time curvature ought somehow to be involved in determining the particles we observe about us. It would seem then, that because of the extreme smallness of numbers, like 2.4×10^{-43}, which govern the size of the gravitational coupling, the gravitational optimist should require that "fantastically strong" gravitational fields must exist (or have existed) somewhere in our universe. By a "fantastically strong" field, I mean one involving radii of curvature (say) as small as the order of an elementary particle size, so that one must expect that the local physics would then be very drastically affected by the presence of gravitation - the local Lorentz covariance having been completely destroyed.

Although the aims of the pure relativist and of the gravitational optimist are so different, there is perhaps a thread of logic connecting them. For if the pure relativist is really a pure mathematician at heart, why does he spend his time with Riemannian manifolds only of dimension four, with that peculiarly arbitrary (+, -, -, -) signature? It is as though some physicist friend of his once told him that such manifolds had to do with the structure of our world - but omitted to mention 10^{-43}! Our pure relativist, having devoted so much of his time to his subject may eventually (in his weaker moments!) feel the need to interest physicists in his results. In this way he may be led to have some sympathy with the views of the gravitational optimist.

The topic I wish to discuss, namely gravitational collapse, illustrates aspects of both the points of view I have mentioned. The role of the pure relativist, here, is in making precise the nature of the situation which arises and in proving exact mathematical theorems applicable to this situation - theorems which one would be hard put to obtain using any other viewpoint. In effect, the pure relativist supplies techniques, whereas it is the gravitational optimist who supplies motivation. The object of the exercise will be, in fact, to establish, on the basis of present physical understanding, the actual existence of regions of enormous curvature in our universe - although the precise results obtained fall perhaps marginally short of this aim.

Let us first establish our conventions. The metric signature (+, -, -, -) will be used, with "gravitational" units chosen, so that

$$G = c = 1 \tag{1}$$

Einstein's equations are then

$$R_{ab} - \tfrac{1}{2}Rg_{ab} + \lambda g_{ab} = -8\pi T_{ab} \tag{2}$$

allowing for the possibility of a cosmological constant λ. I shall not concern myself here with possible modifications of the Einstein theory, except to mention that essentially everything I shall say will apply equally well whether one adopts the standard Einstein theory (2) or the "scalar-tensor" modification of it mentioned by Prof. Dicke in his lecture (Brans-Dicke theory [3, 6]). In the latter case, one must use for the metric ds in what follows, not the most directly physical metric, but the conformally related one in which the main field equations closely resemble Eq. (2) (see Ref. [5]).

I shall consider, first, the case of an exactly spherically-symmetrical gravitational collapse. We know from the classical result of Chandrasekhar [4] (and Landau) that a cold (non-rotating) body which has reached the end-point of nuclear activity and whose mass is larger than about one and one third solar masses cannot hold itself apart against the effects of its own gravitational field. Such a body will therefore collapse and there will be nothing to halt this collapse. Owing to the exact spherical symmetry, the metric exterior to the body must be the Schwarzschild solution [1], which in the usual co-ordinates (taking $\lambda = 0$) is

$$ds^2 = \left(1 - \frac{2m}{r}\right)dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2 (d\theta^2 + \sin^2\theta\, d\varphi^2) \qquad (3)$$

m being the mass of the body in the gravitational units (1). A drawback of the familiar form (3) is that it does not allow us to discuss collapse to within the critical radius given by $r = 2m$. The form clearly becomes singular at $r = 2m$ but this is really a co-ordinate effect arising from the fact that the t-co-ordinate behaves badly at $r = 2m$. We can see this most easily if we replace t by the advanced time parameter v given by

$$V = t + r + 2m \log (r - 2m) \qquad (4)$$

The metric (3) then transforms to the Eddington-Finkelstein form [7, 8]

$$ds^2 = \left(1 - \frac{2m}{r}\right) dv^2 - 2\, dv\, dr - r^2 (d\theta^2 + \sin^2\theta\, d\varphi^2) \qquad (5)$$

The form (5) is non-singular over the entire range $0 < r < \infty$. The vanishing of the coefficient of dv^2 at $r = 2m$ implies merely that the particular co-ordinate hypersurface $r = 2m$ is null. (For finite v, we have $t = \infty$ on $r = 2m$; see Eq. (4).) The space-time Eq. (5) in the neighbourhood of a point at $r = 2m$ is locally Minkowskian, as it also is everywhere else ($r > 0$).

The space-time in the matter region does not have the metric given by Eq. (3) or Eq. (5), but I shall be concerned here only with the empty exterior field. The situation is as depicted in Fig. 1, with one space dimension suppressed. The matter collapses to within $r = 2m$. An observer sitting on the surface of the body would notice nothing peculiar occurring when he crosses $r = 2m$. The curvature (i.e. gravitational tidal forces) would mount steadily, however, becoming finally infinite at $r = 0$. Unlike the hypersurface $r = 2m$ (often misleadingly referred to as the "Schwarzschild singularity" owing to the failure of the t-co-ordinate used in Eq. (3)), the region $r = 0$ is a true physical singularity of the metric and not merely a co-ordinate effect.

Although $r = 2m$ is not a singularity, this region does have some peculiarities from the global point of view. Let us consider an external observer situated some distance from the collapsing body. To ascertain what he sees of the body, we must trace backwards the light rays entering his eye from the direction of the body. From the way the light cones are tipping over near $r = 2m$, it should be evident that, no matter how long he waits, our observer "sees" the body - albeit exceedingly dimly - as it was only just before it plunged into the $r < 2m$ region. In fact, owing to similarities between this and other situations arising in cosmology, we may refer to the $r = 2m$ hypersurface as an event horizon. No signal emitted in the empty $r < 2m$ region can ever escape to the outside, whereas signals emitted in r > 2 m can cross inwards to be trapped inside r = 2m. This "one-way traffic" nature of r = 2 m results from the fact that it is a null hypersurface, i.e. everywhere tangent to the light cone, as is depicted in Fig. 1.

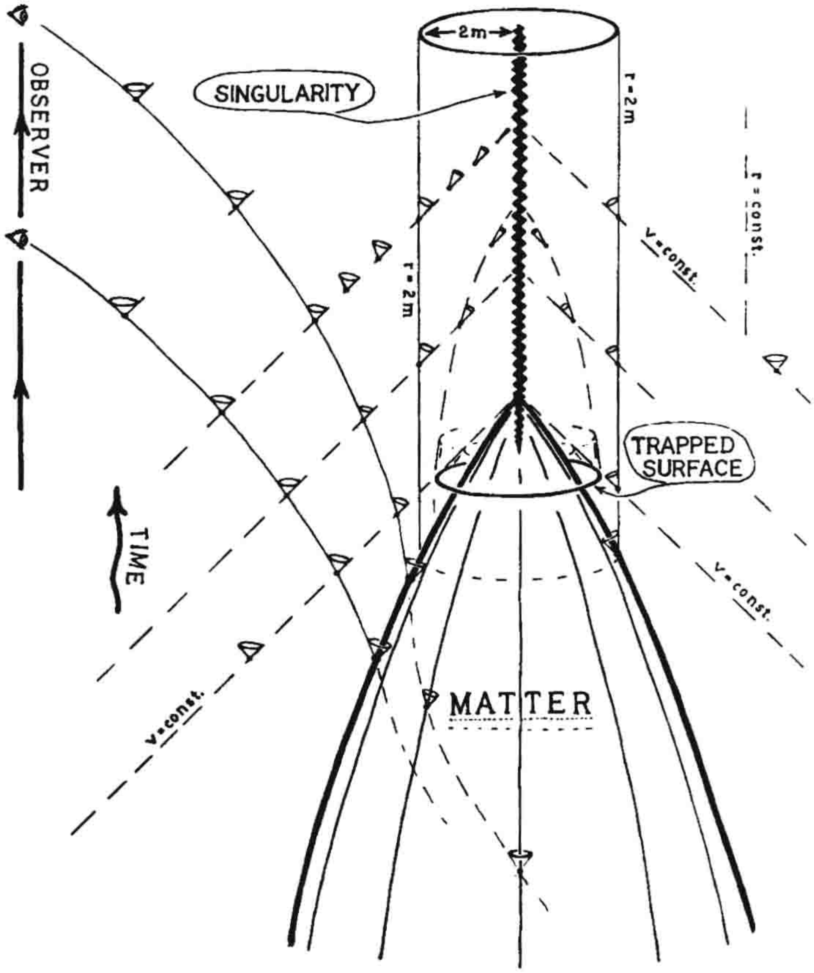


FIG. 1. Space-time picture of a symmetrical collapse to within the Schwarzschild radius r = 2 m.

The radii of space-time curvature in the neighbourhood of r = 2 m (though easily small enough to kill a man in free-fall across r = 2 m in the case of a collapsed star) are not nearly of an order which would impress our gravitational optimist. His attention is drawn instead to the neighbourhood of r = 0, at which the radii of curvature become arbitrarily small. But here we must worry particularly about the implications of our initial assumption of exact spherical symmetry. For it is to be expected that even the slightest departure from spherical symmetry in the external field could result in very gross departures from the metric (5) near r = 0. Thus it is not at all clear that the enormous curvatures desired by our gravitational optimist would result at all in a "generic" physical situation. Indeed, we must even ask whether any analogue of r = 2 m is to be expected in a realistic gravitational collapse. Might not the existence of an event horizon of this type be a very exceptional situation characteristic only of the exact symmetry?

It is in order to answer such questions that the concept of a *trapped surface* is introduced. A trapped surface is a closed (i.e. compact without boundary) spacelike 2-surface $\mathcal{T}$ which has the property that the null geodesics which meet $\mathcal{T}$ orthogonally are all *converging* in the neighbourhood of $\mathcal{T}$. This convergence is measured in the sense of decreasing area of cross-section as we proceed into the future following the null geodesics. In more physical terms, we may envisage a flash of light originating "instantaneously" all over $\mathcal{T}$. This flash spreads away from $\mathcal{T}$ in two directions orthogonal to $\mathcal{T}$ at each point. The characteristic property of a trapped surface is that the area of the flash decreases in *both* directions at every point - that is to say, the *intensity* of the flash everywhere increases. An ordinary spacelike 2-sphere in Minkowski space is clearly not trapped. Here there is an ingoing flash for which the area indeed decreases, but for the outgoing flash the area increases. However, surfaces do exist in Minkowski space which *would* be trapped except for the fact that they are not closed (e.g. the intersection of the past light cones of two spatially separated points). This illustrates the fact that the existence of a trapped surface in a space-time is *not* merely a *local* property of the space-time. For the space-time given by our metric (5) the surfaces r = const., v = const. are trapped for $0 < r < 2m$. Thus, the event horizon $r = 2m$ in fact represents the boundary between the region which contains trapped surfaces and that which does not.

We may picture the situation of Fig. 1 in another way. Let us take a section of Fig. 1 by a spacelike "hyper-plane", that is, $v - r = A$ (= const.). We can think of this as representing the situation at one "instant of time". Superimpose on this picture another section of Fig. 1 taken "a moment later", that is to say by $v - r = A + \epsilon$ ($0 < \epsilon$ = const.). Corresponding to each point P of the "earlier" section, we draw in the intersection of the light cone of P with the "later" section. For example, if, instead of Fig. 1, we had chosen to represent Minkowski space in this way (sections by t = const.), then our resulting picture would consist of a three-dimensional Euclidean space, where each point P would have drawn about it a sphere of radius ϵ and centre P. This sphere and its interior would represent the region that could be causally influenced by an occurrence at P, a time ϵ later; the *surface* of this sphere would represent the location, after time ϵ, of a light flash emitted at P. In the same way, Fig. 2 illustrates the progress of light flashes in the collapse situation of Fig. 1 (taken after the matter has collapsed down to $r = 0$). In Fig. 3 the same situation is depicted on a larger scale, so as to illustrate a trapped surface $\mathcal{T}$ and the locally decreasing surface area of a light flash emitted at $\mathcal{T}$.

Let us now drop the assumption of spherical symmetry - and we may also drop the requirement that our collapsing body be surrounded only by empty space. Can we be *sure* that trapped surfaces can result in a realistic gravitational collapse? I think it is clear that (on the basis of general relativity) trapped surfaces must at least occasionally result from a collapse. The fact that there can be nothing in *principle* against trapped surfaces arising follows from a variety of reasons. The argument I like best is the following "Gedanken-experiment": A mad dictator, who rules a galaxy containing (say) 10^{11} stars, decides that he wishes to

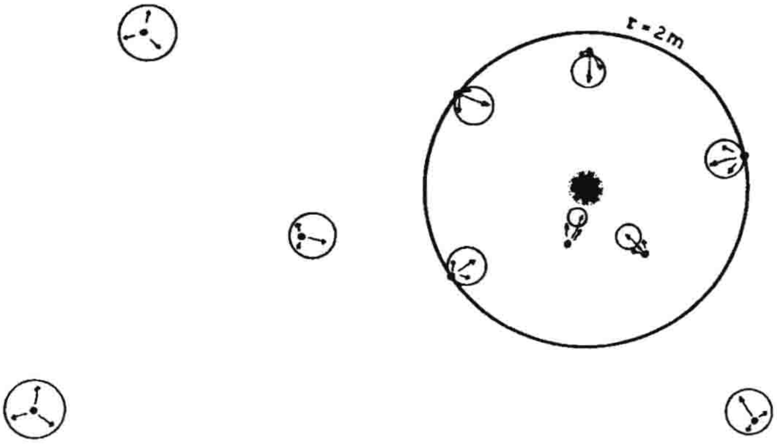


FIG. 2. Space picture of a spherically symmetrical collapsed object.

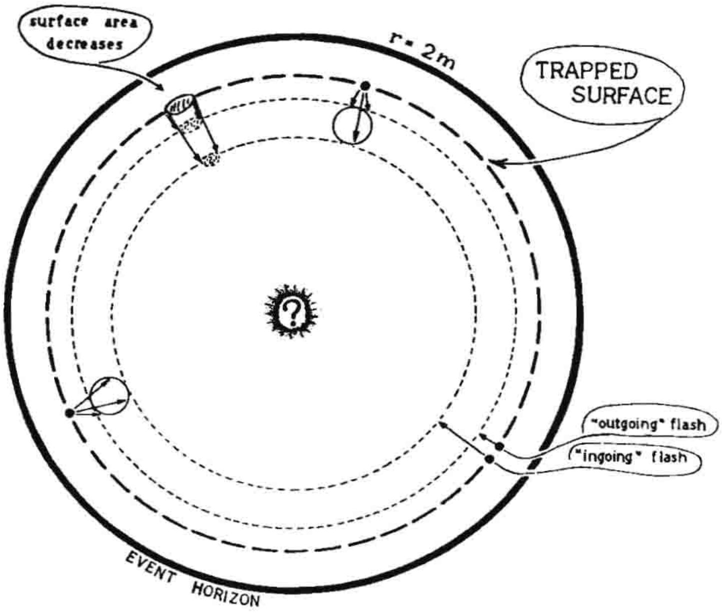


FIG. 3. The locally decreasing surface areas associated with a trapped surface.

build a trapped surface. This he can do (whether or not his galaxy already contains a trapped surface at its centre anyway!) by adjusting the velocities of the stars so that they all fall towards the centre so as to arrive at about the same time. It can then be shown that merely because of the focusing effect on the (idealized) light rays which pass near to (or through) the stars, a trapped surface will result before any serious problem of stellar collisions can arise.

But what are the consequences of a trapped surface arising? In fact, at least two things can be said. In the first instance, it follows from the weak energy assumption (6), which will be discussed shortly, that anything enclosed by a trapped surface (i.e. whose world line threads through it, roughly speaking) cannot be observed by any external observer. Secondly, as a consequence of the theorem to be stated shortly, singularities in space-time must be expected to result.

Before considering the theorem, it is worthwhile to examine, in addition to the Schwarzschild solution, another exact vacuum solution, namely that of Kerr [15] (see also Ref. [2]), since it enables us to incorporate rotation (angular momentum) into the collapse discussion. The Kerr solution depends on two free parameters m and a. Here m describes the total mass (as in the Schwarzschild case) and am the angular momentum. The solution gives the field of only a restricted class of spinning bodies, since the quadrupole and higher moments must all be certain explicit functions of the two parameters a, m. Nevertheless, there are some reasons for believing that whenever a trapped surface develops, the final asymptotic field which results may always be effectively a Kerr solution (or Kerr-Newman solution [16] if there is a non-zero net charge), all the excess multipole moments being ultimately radiated away by gravitational (plus electromagnetic) radiation, as the field itself finally settles down after the collapse has taken place. This is largely conjecture at present, but appears to be borne out in the non-rotating case at least, by a result due to Israel [14].

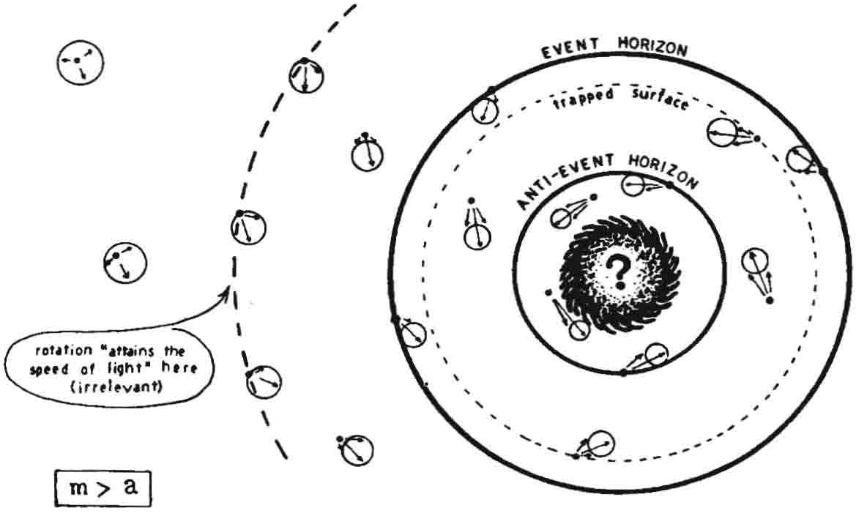


FIG. 4. Space picture of a Kerr solution for "small" angular momentum, representing a rotating collapsed object.

There are two cases to consider in the Kerr solution. First there is the case m > a of small angular momentum. The situation is depicted in Fig. 4 (viewed along the rotation axis), which is drawn according to the same conventions as Figs 2 and 3. Note that external to the event horizon there is a hypersurface (irrelevant for our purposes) inside which the system is "rotating faster than light", that is to say, a particle would have to exceed the local light velocity in order to appear to be stationary to an observer at infinity. (The picture unduly emphasizes the importance of this region - our pure relativist would not much care for Figs 2 - 5!) The event horizon here is a stationary null hypersurface. Signals can cross it from the outside to the inside but not the other way about. There is also another stationary null hypersurface inside the event horizon which we may refer to as the *anti-event horizon*. There is an inward directed one-way traffic of signals between the event horizon and the anti-event horizon. But inside the anti-event horizon signals can pass freely both inwards and outwards provided they spiral around sufficiently. Trapped surfaces reside in the region between the event horizon and the anti-event horizon. This region acts as a kind of *shield* for whatever chaos resides inside.

Now consider a sequence of Kerr solutions for which m is kept fixed but a is steadily increased. The anti-event horizon gets nearer and nearer to the event horizon until finally they annihilate one another leaving the internal chaos visible from the outside. However, it should be emphasized that for any *particular* collapse situation in which trapped surfaces do form, the event horizon can never be annihilated in this way. The internal chaos must be forever hidden from the outside world. If a collapse situation gives rise to a metric similar to a Kerr solution with a > m, then it cannot do so via metrics containing trapped surfaces. Furthermore, since we have no trapped surface for a > m, we also have no theorem predicting enormous curvature regions - at least, this is the situation at present.

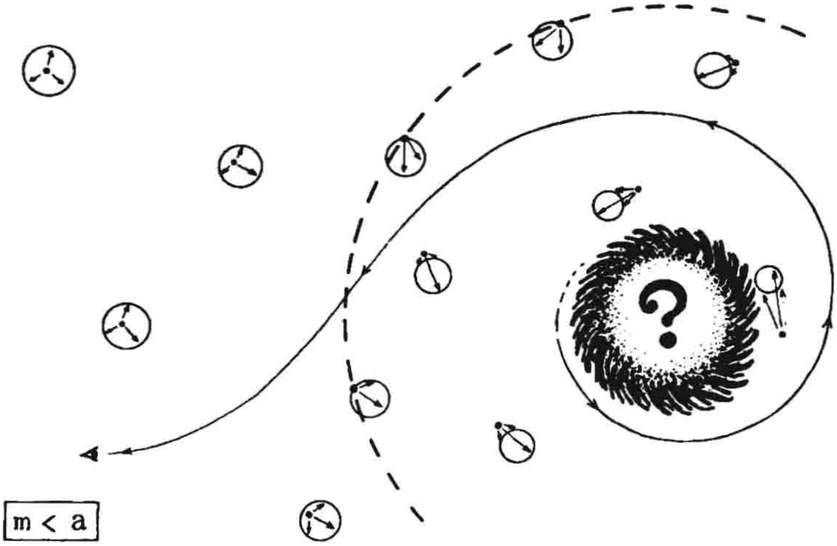


FIG. 5. Space picture of a Kerr solution for "large" angular momentum. Signals can escape from the internal region to be observed by an external observer.

Thus, the question mark at the centre of Fig.5 is really a different kind of question mark from that in Fig.4. In a *generic* (physically realistic) situation like Fig.4, we know that there still remains a region that we cannot understand on the basis of present space-time theory (a "singularity"), while for a generic Fig.5 we do not know whether there is such a region. A point of some interest, but whose significance is not clear, is that the singularity region of Fig.4 in the *exact* Kerr solution resides some way away from the event horizons and is qualitatively *not* different from the singularity region of Fig.5. Thus it might seem that the nature of the true singularity (e.g. enormous curvature region) is not sensitive to the presence or absence of an event horizon[1]. This may be misleading, however, as the existing singularity theorems give no indication as to the location of the singularities in generic situations and very little indication as to their nature. I think that the best description of a generic collapse with a > m (if such occurs) is, at present, simply a big question mark.

I now wish to describe a precise mathematical theorem which forms the basis of many of the above remarks concerning the existence of singularity regions. Since we have no desire to restrict attention just to vacuum solutions nor just to matter with some particular equation of state, the content of Einstein's equations (2) will manifest itself only as inequalities on the Ricci curvature. Such inequalities we can refer to as energy conditions. The two energy conditions of interest to us here are the following:

The weak energy condition:

$$\ell^a \ell_a = 0 \text{ implies } R_{ab}\ell^a \ell^b \leqslant 0 \tag{6}$$

The strong energy condition:

$$t^a t_a = 1 \text{ implies } R_{ab}\, t^a t^b \leqslant 0 \tag{7}$$

(That condition (6) is a consequence of condition (7) follows by a limiting argument.)

The interpretation of the "weak" condition (6) is that with Einstein's equations (2) it becomes

$$E + p_i \geqslant 0 \qquad (i = 1, 2, 3) \tag{8}$$

where in an eigentetrad of T_{ab}, E is the energy density and p_1, p_2, p_3 are the three principal pressures (unequal only for anisotropic media). In fact Eq.(6) (or Eq.(8)) is a consequence of the non-negative definiteness of the (local) energy density (that is: $t^a t_a = 1$ implies $T_{ab}\ t^a t^b \geqslant 0$). Thus, we expect any reasonable matter to satisfy Eq.(8). From this will follow all the essential properties of trapped surfaces mentioned earlier, namely that space-time singularities must be expected to result, in a region which is not visible from the outside. However, a loophole remains in the application of the precise theorem [17, 18] which uses only Eq.(6). (This is the possibility of the non-existence of an open global Canchy hypersurface in our universe). This loophole can be removed if we adopt Eq.(7) rather than Eq.(6).

---

[1] I have been somewhat vague about the meaning of the term "event horizon" for a generic situation. More accurately, we could call this the "common event horizon for all external inertial observers". It represents the boundary of the set of all points which cannot be connected to external infinity by time-like future-directed curves (assuming asymptotic flatness). This set contains all the trapped surfaces, but will in general extend out somewhat beyond the trapped surface region.

The interpretation of the "strong" condition (7), provided $\lambda \leq 0$ in Einstein's equations (2), is

$$E + \sum p_i \geqslant 0 \ \ \underline{\text{and}} \ \ E + p_i \geqslant 0 \qquad (i = 1, 2, 3) \tag{9}$$

Although this does not quite follow from the non-negative definiteness of energy, it is still a very reasonable assumption for "normal" matter. (Equivalent to Eq.(9) is: $t^a t_a = 1$ implies $T_{ab}\, t^a t_b > \frac{1}{2} T^c_c$.) Much use has been made of Eq.(7) by Hawking, who has obtained several important theorems [9 - 11] predicting singularities, mainly in cosmological contexts. He has referred, descriptively, to Eq.(7) as stating that "gravitation is always attractive". A slight drawback to Eq.(7) is that its strict applicability requires (unlike Eq.(6)) that the cosmological constant λ be zero or negative. However, in regions where curvature is already large compared with λ, it is difficult to believe that the presence of a cosmological constant should seriously affect the subsequent discussion.

The theorem I wish to describe was obtained jointly with Hawking [13]. In its main implications it incorporates and somewhat generalizes the previous results. Ideally, one would like to be able to show that on the basis of Einstein's equations and reasonable physical assumptions, the space-time curvature becomes arbitrarily large (in an appropriate sense) in some region. Unfortunately none of the theorems to date directly shows this. Instead, we must content ourselves with a deduction that the space-time in question is not geodesically complete and, consequently, not extendible to a geodesically complete space-time. By a space-time, I shall here mean a pseudo-Riemannian four-manifold with a (+, -, -, -) signature, which is time-orientable. Such a manifold is timelike- and null-geodesically complete if every timelike or null geodesic can be extended to arbitrarily large values of its affine parameter. Physically, this means (crudely) that freely moving particles or photons cannot "just disappear" or "just appear" in the space-time.

Theorem: Space-time cannot be timelike- or null-geodesically complete if

(i) the strong energy condition holds

(ii) there are no closed timelike curves

(iii) every timelike or null geodesic (with tangent vector t^a) encounters a region where

$$t_{[a} R_{b]cd[e}\, t_{f]}\, t^c t^d \neq 0$$

(iv) there exists a compact achronal set with a trapped edge.

Assumption (ii) is the physically very reasonable requirement that it should be impossible for any observer to travel (or to send signals) into his own past. Assumption (iii) would be expected to hold in any physically realistic or "generic" space-time. Randomly oriented curvature (e.g. gravitational waves, no matter how weak) would be sufficient to ensure that assumption (iii) holds.

The physically important assumption of the theorem is (iv). An achronal set (sometimes referred to as a semi-spacelike set [18]) is a set of points no two of which can be connected by a strictly timelike curve (e.g. a single point or, in normal circumstances, a spacelike or null hypersurface). The "edge" of such a set S is "trapped" if the null geodesics, lying on the boundary of the future of S, leave S at points at which they locally converge into the future (as with a trapped surface). It is not necessary here to appreciate the meaning of assumption (iv) in its full generality. Only the following three special cases will concern us, any one of which can be inserted in the theorem in place of (iv):

(iva) there exists a trapped surface
(ivb) there exists a compact spacelike hypersurface
(ivc) all the past-directed null geodesics through some point start converging again.

If we use (iva), we effectively obtain the required implication that, on the basis of Einstein's equations, trapped surfaces result in singularities in space-time. Alternatively we can use (ivb), which is essentially the statement that the universe should be *spatially closed*. Again the implication is that space-time singularities should occur. The final alternative (ivc) has relevance to observational questions. The point referred to can be taken to represent the earth at the present moment. The past-directed null geodesics represent the portion of the universe that we can "see". Roughly speaking, (ivc) states that the solid angle subtended by an object of given size will have a minimum value. Hawking and Ellis [12] have suggested that the present observations of the 3°K "black-body" radiation indicate that (ivc) is satisfied in our universe. I believe Sciama discusses this in greater detail in his lecture [19]. Again, space-time singularities (this time in the past) are to be inferred.

I cannot go into the proof of the theorem here. The methods used involve differential topology; also certain arguments rather specific to the study of space-times have had to be developed. These are very much the realm of the pure relativist.

It is one of the ironies of the subject[2] that by obtaining results of this kind, our pure relativist pulls the rug from beneath him! Now he can no longer regard his space-time as something remote from the rest of physics. And if he is to have any hope at all of appreciating the nature of one of his "singularities", or of the neighbouring space-time, he will presumably have to descend his ivory tower into the murky depths of quantum theory and particle physics - and even this may well not be sufficient!

REFERENCES

[1] BIRKHOFF, G.D., Relativity and Modern Physics, Harvard Univ. Press, Cambridge, Mass. (1923).
[2] BOYER, R.H., LINDQUIST, R.W., J. math. Phys. 8 2 (1967) 265.
[3] BRANS, C., DICKE, R.H., Phys. Rev. 124 (1961) 925.
[4] CHANDRASEKHAR, S., R. astr. Soc. monthly Notices 95 (1935) 207.
[5] DICKE, R.H., Phys. Rev. 125 (1962) 2163.

---

[2] There is, perhaps, a twin irony that our gravitational optimist cannot "see" (and therefore cannot "use") the singularity that he had commissioned the pure relativist to find for him inside the collapsing star. But I do not feel that this is quite valid, since the gravitational optimist has also been presented with a "visible" singularity in the remote past.

[6] DICKE, R.H., these Proceedings.
[7] EDDINGTON, A.S., Nature 113 (1924) 192.
[8] FINKELSTEIN, D., Phys. Rev. 110 (1958) 965.
[9] HAWKING, S.W., Proc. R. Soc. A294 (1966) 511.
[10] HAWKING, S.W., Proc. R. Soc. A295 (1966) 490.
[11] HAWKING, S.W., Proc. R. Soc. A300 (1967) 187.
[12] HAWKING, S.W., ELLIS, G.F.R., Astrophys. J. 152 (1968) 25.
[13] HAWKING, S.W., PENROSE, R., to be published.
[14] ISRAEL, W., Phys. Rev. 164 (1967) 1776.
[15] KERR, R.P., Phys. Rev. Lett. 11 (1963) 237.
[16] NEWMAN, E.T., COUCH, E., CHINNAPARED, K., EXTON, A., PRAKASH, A., TORRENCE, R., J. math. Phys. 6 (1965) 918.
[17] PENROSE, R., Phys. Rev. Lett. 14 (1965) 57.
[18] PENROSE, R., "Structure of space-time", in Battelle Rencontres, Seattle, 1967 (De WITT, C., WHEELER, J.A., Eds), Benjamin, New York (1968).
[19] SCIAMA, D.W., these Proceedings.

39

SOLUTIONS OF THE ZERO-REST-MASS EQUATIONS

J. Math. Phys. **10(1)** (1969), 38–9

This paper, displaying the twistor contour-integral formula in a simple (not visibly twistorial) form, was received in September 1966, being for some reason delayed (lost?) for about two years and four months, before it was published. But I suppose I should not complain, because, unbeknown to me, Harry Bateman had already found a formula equivalent to mine for the scalar case in 1904 (Bateman, H., 1904, The solution of partial differential equations by means of definite integrals. *Proc. Lond. Math. Soc.* **(2)1**, 451–8) and he had also found a corresponding expression for the Maxwell case which he published in his 1944 book, *Partial Differential Equations of Mathematical Physics* (Dover, New York). Bateman's formulae were not explicitly expressed in terms of contour integration, however. I have recently been informed that E. T. Whittaker had a formula of this nature, for general solutions of Maxwell's equations, already in 1903 (Whittaker, E. T. (1903) On the partial differential equations of mathematical physics. *Math. Ann.* **57**, 333–55) from which Bateman's formulae was apparently developed.

JOURNAL OF MATHEMATICAL PHYSICS VOLUME 10, NUMBER 1 JANUARY 1969

Solutions of the Zero-Rest-Mass Equations

ROGER PENROSE
Birkbeck College, London, England

(Received 23 September 1966)

By means of contour integrals involving arbitrary analytic functions, general solutions of the zero-rest-mass field equations in flat space–time can be generated for each spin. If the contour surrounds only a simple (respectively, low-order) pole of the function, the resulting field is null (respectively, algebraically special).

1. THE CONTOUR INTEGRAL

It is possible to generate a very wide class of solutions to the zero-rest-mass field equations for each spin, $s = 0, \frac{1}{2}, 1, \frac{3}{2}, 2, \cdots$, in flat space–time by means of a certain contour-integral expression. By choosing the integrand and contour suitably, the resulting field can be made to have certain prescribed properties, e.g., it may be made null or algebraically special. The expression arises naturally in the theory of twistors,[1] but the result can be given quite readily without using twistors. The present note gives the main results without going into the general theory.

Let x^0, x^1, x^2, x^3 be standard Minkowskian coordinates and set

$$u = 2^{-\frac{1}{2}}(x^0 + x^1), \quad v = 2^{-\frac{1}{2}}(x^0 - x^1), \quad \zeta = 2^{-\frac{1}{2}}(x^2 + ix^3), \tag{1.1}$$

so that the metric becomes $ds^2 = 2du\,dv - 2d\zeta\,d\bar{\zeta}$. Let f be an analytic function of three complex variables. Choose a nonnegative integer $2s$ and, for $r = 0, 1, \cdots, 2s$, put

$$\phi_r = \frac{1}{2\pi i}\oint \lambda^r f(\lambda, u + \lambda\bar{\zeta}, \zeta + \lambda v)\,d\lambda, \tag{1.2}$$

where the contour surrounds (but does not encounter) singularities of f and varies continuously with u, v, and ζ. Then we have

$$\frac{\partial\phi_r}{\partial\bar{\zeta}} = \frac{\partial\phi_{r+1}}{\partial u}, \frac{\partial\phi_r}{\partial v} = \frac{\partial\phi_{r+1}}{\partial\zeta}, \quad r = 0, \cdots, 2s - 1, \tag{1.3}$$

(if $s > 0$) and

$$\left\{\frac{\partial^2}{\partial u\partial v} - \frac{\partial^2}{\partial\zeta\partial\bar{\zeta}}\right\}\phi_r = 0, \quad r = 0, \cdots, 2s. \tag{1.4}$$

Equations (1.4) is simply the wave equation in the coordinates (1.1) while Eqs. (1.3) are the spin-s zero-rest-mass equations in a suitable notation. For if we put

$$\phi_0 = \phi_{000\cdots0}, \quad \phi_1 = \phi_{100\cdots0}, \quad \phi_2 = \phi_{110\cdots0}, \cdots, \quad \phi_{2s} = \phi_{111\cdots1}, \tag{1.5}$$

where $\phi_{ABC\cdots K}$ has $2s$ indices and is symmetric: $\phi_{AB\cdots K} = \phi_{(AB\cdots K)}$; then Eqs. (1.3) can be written

$$\frac{\partial}{\partial x_{AP'}}\phi_{ABC\cdots K} = 0 \tag{1.6}$$

(summation convention assumed), where the 2-spinor notation $x_{00'} = v$, $x_{01'} = -\bar{\zeta}$, $x_{10'} = -\zeta$, and $x_{11'} = u$ is being used. Equation (1.6) is simply the Dirac–Fierz spinor equation[2,3] for mass zero and spin s. If $s = 1$, we can put

$$\phi_0 = \tfrac{1}{2}(F^{12} - F^{02} - iF^{13} + iF^{03}),$$
$$\phi_1 = \tfrac{1}{2}(F^{01} - iF^{23}),$$
$$\phi_2 = \tfrac{1}{2}(F^{12} + F^{02} + iF^{13} + iF^{03}),$$

and Eqs. (1.3) become Maxwell's equations

$$\frac{\partial}{\partial x^a}F^{ab} = 0, \quad \frac{\partial}{\partial x^a}F_{bc} + \frac{\partial}{\partial x^b}F_{ca} + \frac{\partial}{\partial x^c}F_{ab} = 0.$$

Similarly, for $s = 2$, we can get the linearized Einstein equations in gauge-invariant ("curvature-tensor") form.[3,4,5]

2. NULL AND ALGEBRAICALLY SPECIAL FIELDS

Suppose the contour in (1.2) surrounds only a k-order pole of the function f. Let $\lambda = \eta$ be the pole for given u, v, ζ (so η is a function of u, v, ζ, $\bar{\zeta}$). Then

$$\oint (\lambda - \eta)^k \lambda^r f(\lambda, u + \lambda\bar{\zeta}, \zeta + \lambda v)\,d\lambda = 0.$$

Thus, if $2s \geq k$,

$$\phi_{r+k} - k\phi_{r+k-1}\eta + \tfrac{1}{2}k(k - 1)\phi_{r+k-2}\eta^2 - \cdots + \phi_r(-\eta)^k = 0, \quad r = 0, \cdots, 2s - k \tag{2.1}$$

[see (1.2)]. We can rewrite Eqs. (2.1), using the notation (1.5), as

$$\phi_{AB\cdots DE\cdots K}\xi^A\xi^B\cdots\xi^D = 0, \tag{2.2}$$

where $A, B, \cdots, D$ are k in number and where

$$\xi^0 = -\eta, \quad \xi^1 = 1. \tag{2.3}$$

---

[1] R. Penrose, J. Math. Phys. **8**, 345 (1967).

[2] P. A. M. Dirac, Proc. Roy. Soc. (London) **A155**, 447 (1936).

[3] M. Fierz, Helv. Phys. Acta **13**, 45 (1940); M. Fierz and W. Pauli, Proc. Roy. Soc. (London) **A173**, 211 (1939).

[4] R. K. Sachs and P. G. Bergmann, Phys. Rev. **112**, 674 (1958).

[5] R. Penrose, Proc. Roy. Soc. (London) **A284**, 159 (1965); Ann. Phys. (N.Y.) **10**, 171 (1960).

Equation (2.2) is, in fact, precisely the condition for a spinor ξ^A to represent a $(2s - k + 1)$-fold *principal null direction*[5] for the field $\phi_{AB\cdots K}$. Thus, when $k = 1$, so that the contour surrounds only a simple pole of f, all $2s$ principal null directions coincide in the direction of ξ^A so that we get a *null field.*[5] [In the Maxwell case this means $F_{ab}F^{ab} = 0 = F_{ab}F_{cd}\epsilon^{abcd}$ i.e., $\phi_0\phi_2 = (\phi_1)^2$.] More generally, whenever $k < 2s$ *at least two* principal null directions coincide (in the direction of ξ^A), that is to say, the field is algebraically special. (Such fields, when $s = 2$, are of interest in gravitation theory.[6,7])

The direction of ξ^A is given by the vector translation $\xi^A\bar{\xi}^{B'}$, i.e., by

$$\begin{aligned} du:d\zeta:d\bar{\zeta}:dv &= \xi^0\bar{\xi}^{0'}:\xi^0\bar{\xi}^{1'}:\xi^1\bar{\xi}^{0'}:\ \xi^1\bar{\xi}^1 \\ &= \eta\bar{\eta}:-\eta:-\bar{\eta}:1. \end{aligned}$$

Thus η defines the null direction[8] given by

$$du + \eta\, d\bar{\zeta} = 0 = d\zeta + \eta\, dv. \tag{2.4}$$

It is known[5–7,9] that the multiple principal null direction of an algebraically special field is tangential to a shear-free congruence of null geodesics (here, straight lines). In the above case, this follows also by a theorem of Kerr[1] which states that such a congruence is defined by (2.4) if we *specify an analytic relation connecting* η, $u + \eta\bar{\zeta}$, and $\zeta + \eta v$. In the present situation, the analytic equation $\{f(\lambda, u + \lambda\bar{\zeta}, \zeta + \lambda v)\}^{-1} = 0$ defines the poles $\lambda = \eta$ of f, verifying that the directions (2.4) are indeed geodetic and shear free.

In addition, Kerr's theorem states that, conversely, *any* shear-free geodetic null congruence in flat space–time (except for certain rather special limiting cases) can be obtained from such a analytic relation. This indicates the generality of expression (1.2) for the construction of null fields. Robinson[9] showed how, *starting* from any shear-free geodetic null congruence, it is possible to construct all the corresponding null solutions of Maxwell's equations by the arbitrary specification of an analytic function of two complex variables. The integral (1.2) achieves effectively the same thing (in flat space–time, but now for fields of arbitrary spin $s \geq 1$). For, by Kerr's result, the given shear-free congruence can (normally) be defined by (2.4) subject to $h(\eta, u + \eta\xi, \zeta + \eta v) = 0$, where h is some analytic function with simple zeros. Into the integrand of (1.2) we can substitute $f = gh^{-1}$, where g is an analytic function regular at the (relevant) zeros of h. The freedom of choice for the residues in (1.2) at the poles of f is, h being given, simply the freedom in the choice of g at $h = 0$. This is essentially one complex function of two complex variables ($h = 0$ being a two-complex-dimensional set) in agreement with Robinson's result.

Indeed, we can go somewhat further since algebraically special fields can also be treated by the method given here. For example, if $s = 2$, an expression gh^{-3} yields, when substituted into (1.2), a general type of algebraically special linearized gravitational field. (Here it is the values of g and its first and second derivatives, at $h = 0$, which are relevant.) We can also consider the slightly more general fields given when $k = 2s$ in (1.2). The directions (2.4) are evidently (by Kerr's theorem) still geodetic and shear free but they are now just *simple* principal null directions and the field is not algebraically special. Since, for a general field, the principal null directions are neither shear free nor geodetic, it follows that the fields given by $k = 2s$ in (1.2) are still of a rather special type. (The case $k = 2s = 1$ defines what we might tentatively call a "null neutrino field.") However, more general fields are generated if the contour surrounds a pole of higher order than $2s$ [for then (2.4) will not even be a principal null direction of the field], or more than one pole of f (in which case the resulting field will be a finite linear combination of fields of the type we have just been considering), or singularities or singular regions of more complicated types.

It is not hard to construct a function f for most of the simple types of fields normally encountered (e.g., plane waves, monopole, or multipole solutions, etc.). Also, provided the contour can be chosen consistently, we can obtain linear combinations of such fields in the form (1.2) simply by taking the corresponding linear combinations of f's. This process may fail if too extensive (continuous) linear combinations of f's are taken, since the resulting singularities may leave no room for the contour. Nevertheless, it is evident that there is considerable generality in the expression (1.2).

The full discussion of (1.2) and of its transformation properties is best carried out in terms of twistors.[1] The twistor description will be given elsewhere.[10]

---

[6] I. Robinson and A. Trautman, Proc. Roy. Soc. (London) **A265**, 463 (1962).

[7] J. N. Goldberg and R. K. Sachs, Acta Phys. Polon. **22**, 13 (1962).

[8] The value $\eta = \infty$ also gives a well-defined null direction, although this would not arise from the integral (1.2) as given. To obtain null and algebraically special fields in a way similar to the above, but in which this exceptional null direction could also be represented, we would have to transform (1.2) suitably. It is in the transformation properties of (1.2) that the different spin values play a role. [Equation (1.2) is curiously oblivious to the value of s here!] The complete manifestly (conformally) covariant expression, of which (1.2) is a particular realization, requires the use of twistors.

[9] I. Robinson, J. Math. Phys. **2**, 290 (1960).

[10] *Note Added in Proof.* Due to the delay in the publishing of this paper, this description has already appeared; see R. Penrose, Intern. J. Theoret. Phys. **1**, 61 (1968).

40

GRAVITATIONAL COLLAPSE: THE ROLE OF GENERAL RELATIVITY

Rivista del Nuovo Cimento, Serie 1, Vol 1, Numero Speciale, 252–276

There are many articles in these *Collected Works* concerned with gravitational collapse to a black hole. It may well be appropriate to single out this particular one, for the number of issues it raises that had perhaps not been properly appreciated before. Among these is the question of *cosmic censorship,* according to which naked singularities are claimed not to arise in a generic gravitational collapse, so that an actual black hole is necessarily formed. Another conjecture is pointed out—referred to here as 'GIC' (generalized Israel conjecture)—which asserts that a black hole, when left to settle down, will become a Kerr space-time. Unlike cosmic censorship, which remains unproved, this latter conjecture is now essentially a theorem, through the work of Werner Israel, Brandon Carter, David Robinson, Stephen Hawking, and others (the result being now commonly referred to as the 'no hair theorem'). This article also points out how rotational energy can in principle be extracted from a black hole, using a key property of its Killing vectors which is now regarded as being central to the proposed mechanisms whereby quasars extract vast amounts of energy from central black holes.

Gravitational collapse: The role of general relativity. *Rivista del Nuovo Cimento,* Serie 1, Vol 1, Numero Speciale, 252–276

Gravitational Collapse: the Role of General Relativity.

R. PENROSE

Department of Mathematics, Birkbeck College - London

Stars whose masses are of the same order as that of the sun ($M_\odot$) can find a final equilibrium state either as a white dwarf or, apparently, (after collapse and ejection of material) as a neutron star. These matters have been nicely discussed in the lectures of HEWISH and SALPETER. But, as they have pointed out, for larger masses no such equilibrium state appears to be possible. Indeed, many stars are observed to have masses which are much larger than $M_\odot$—so large that it seems exceedingly unlikely that they can ever shed sufficient material so as to be able to fall below the limit required for a stable white dwarf ($\sim 1.3\, M_\odot$: CHANDRASEKHAR [1]) or neutron star ($\sim 0.7\, M_\odot$: OPPENHEIMER and VOLKOFF [2]) to develop. We are thus driven to consider the consequences of a situation in which a star collapses right down to a state in which the effects of general relativity become so important that they eventually dominate over all other forces.

I shall begin with what I think we may now call the « classical » collapse picture as presented by general relativity. Objections and modifications to this picture will be considered afterwards. The main discussion is based on Schwarzschild's solution of the Einstein vacuum equations. This solution represents the gravitational field exterior to a spherically symmetrical body. In the original Schwarzschild co-ordinates, the metric takes the familiar form

$$\text{(1)} \qquad \mathrm{d}s^2 = (1 - 2m/r)\, \mathrm{d}t^2 - (1 - 2m/r)^{-1}\, \mathrm{d}r^2 - r^2(\mathrm{d}\theta^2 + \sin^2\theta\, \mathrm{d}\varphi^2)\,.$$

Here θ and φ are the usual spherical polar angular co-ordinates. The radial co-ordinate r has been chosen so that each sphere $r = \text{const}$, $t = \text{const}$ has intrinsic surface area $4\pi r^2$. The choice of time co-ordinate t is such that the metric form is invariant under $t \to t + \text{const}$ and also under $t \to -t$. The static nature of the space-time is thus made manifest in the formal expression for the metric. The quantity m is the mass of the body, where « general-relativistic units » are chosen, so that

$$c = G = 1\,,$$

that is to say, we translate our units according to

$$1\ \mathrm{s} = 3\cdot 10^{10}\ \mathrm{cm} = 4\cdot 10^{38}\ \mathrm{g}\,.$$

When $r = 2m$, the metric form (1) breaks down. The radius $r = 2m$ is referred to as the *Schwarzschild radius* of the body.

Let us imagine a situation in which the collapse of a spherically symmetrical (nonrotating) star takes place and continues until the surface of the star approaches the Schwarzschild radius. So long as the star remains spherically symmetrical, its *external field* remains that given by the Schwarzschild metric (1). The situation is depicted in Fig. 1. Now the particles at the surface of the star must describe *timelike* lines. Thus, from the way that the « angle » of the light cones appears to be narrowing down near $r = 2m$, it would seem that the surface of the star can never cross to within the $r = 2m$ region. However, this is misleading. For suppose an observer were to follow the surface of the star in a rocket ship, down to $r = 2m$. He would find (assuming that the collapse does not differ significantly from free fall) that the total proper time that he would experience as elapsing, as he finds his way down to $r = 2m$, is in fact *finite*. This is despite the fact that the world line he follows has the *appearance* of an « infinite » line in Fig. 1. But what does the observer experience after this finite proper time has elapsed? Two possibilities which suggest themselves are: i) the observer encounters some form of space-time singularity —such as infinite tidal forces—which inevitably destroys him as he approaches $r = 2m$; ii) the observer enters some region of space-time not covered by the (t, r, θ, φ) co-ordinate system used in (1). (It would be unreasonable to suppose that the observer's experiences could simply *cease* after some finite time, without his encountering some form of violent agency.)

In the present situation, in fact, it is possibility ii) which occurs. The easiest way to see this is to replace the co-ordinate t by an advanced time parameter v given by

$$v = t + r + 2m\log(r - 2m)\,,$$

whereby the metric (1) is transformed to the form (EDDINGTON [3], FINKELSTEIN [4])

$$ds^2 = (1 - 2m/r)\,dv^2 - 2\,dr\,dv - r^2(d\theta^2 + \sin^2\theta\,d\varphi^2)\,. \tag{2}$$

CHAPTER 40

This form of metric has the advantage that it does not become inapplicable at $r = 2m$. The whole range $0 < r < \infty$ is encompassed in a nonsingular fashion by (2). The part $r > 2m$ agrees with the part $r > 2m$ of the original expression (1). But now the region has been extended inwards in a perfectly regular way across $r = 2m$ and right down towards $r = 0$.

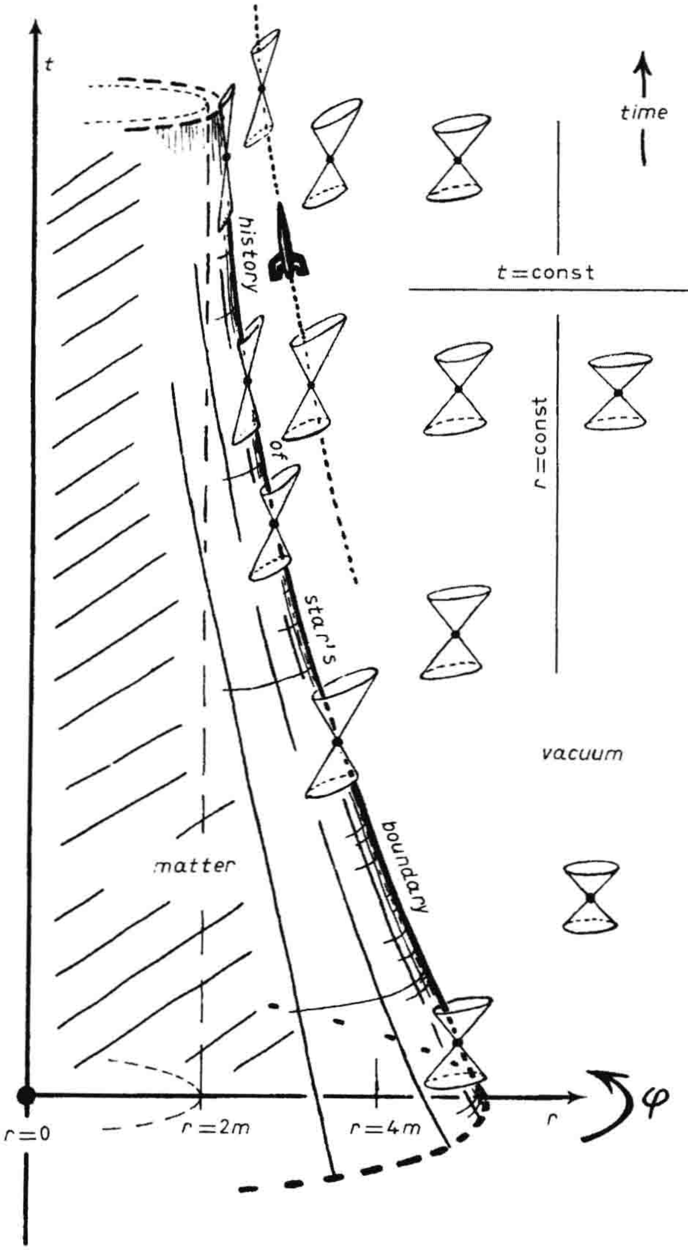

t
time
t=const
r=const
history
of
star's
boundary
vacuum
matter
r=0
r=2m
r=4m
r
φ

Fig. 1. – Spherically symmetrical collapse in the usual Schwarzschild co-ordinates

The situation is as depicted in Fig. 2. The light cones tip over more and more as we approach the centre. In a sense we can say that the gravitational field has become so strong, within $r = 2m$, that even light cannot escape and is dragged inwards towards the centre. The observer on the rocket ship, whom we considered above, crosses freely from the $r > 2m$ region into the $0 < r < 2m$ region. He encounters $r = 2m$ at a perfectly finite time, according to his own local clock, and he experiences nothing special at that point. The space-time there is locally Minkowskian, just as it is everywhere else ($r > 0$).

Let us consider another observer, however, who is situated far from the star. As we trace the light rays from his eye, back into the past towards the star, we find that they cannot cross into the $r < 2m$ region after the star has collapsed through. They can only intersect the star at a time *before* the star's surface crosses $r = 2m$. No matter how long the external observer waits, he can always (in principle) still see the surface of the star as it *was* just before it plunged through the Schwarzschild radius. In practice, however, he would soon see nothing of the star's surface—only a « black hole »—since the observed intensity would die off exponentially, owing to an infinite red shift.

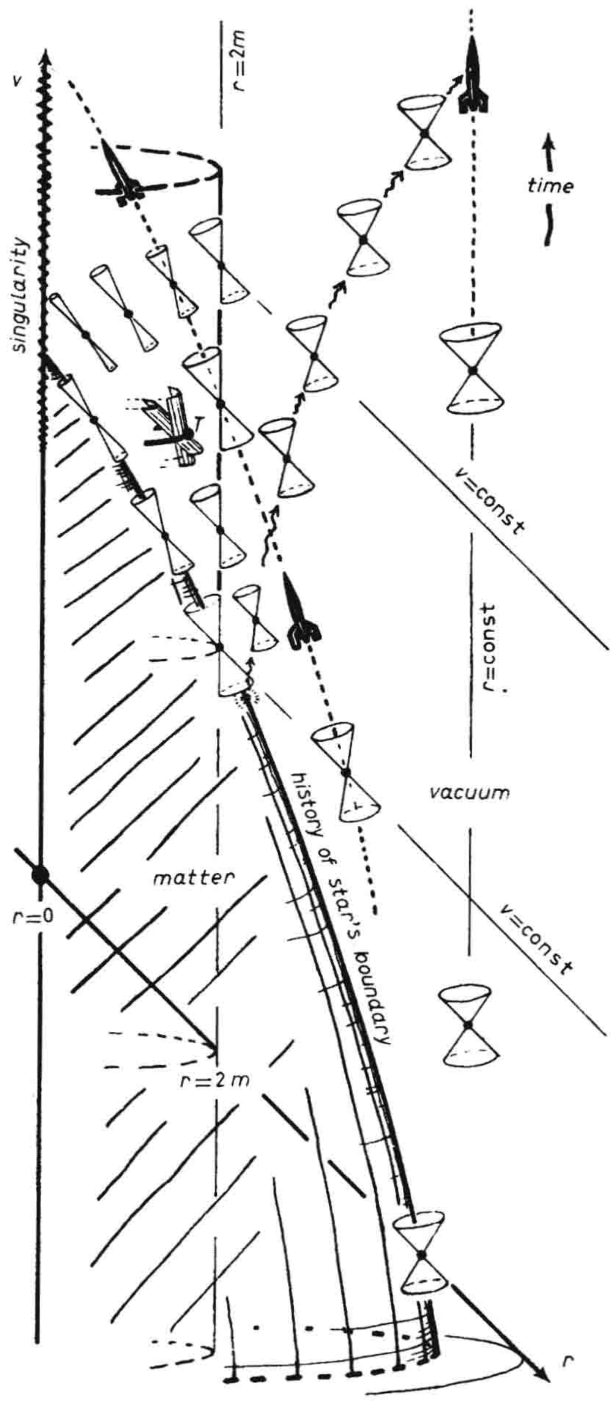


Fig. 2. – Spherically symmetrical collapse in Eddington-Finkelstein co-ordinates.

But what will be the fate of our original oberver on the rocket ship? After crossing the Schwarzschild radius, he finds that he is compelled to enter regions of smaller and smaller r. This is clear from the way the light cones tip over towards $r = 0$ in Fig. 2, since the observer's world line must always remain a timelike line. As r decreases, the space-time curvature mounts (in proportion to r^{-3}), becoming theoretically infinite at $r = 0$. The physical effect of space-time curvature is experienced as a *tidal force*: objects become squashed in one direction and stretched in another. As this tidal effect mounts to infinity, our observer must eventually (*) be torn to pieces—indeed, the very atoms of which he is composed must ultimately individually share this same fate!

Thus, the true *space-time singularity*, resulting from a spherically symmetrical collapse, is located not at $r = 2m$, but at $r = 0$. Although the hypersurface $r = 2m$ has, in the past, itself been frequently referred to as the « Schwarzschild singularity », this is really a misleading terminology since $r = 2m$ is a singularity merely of the t co-ordinate used in (1) and not of the space-time geometry. More appropriate is the term « event horizon », since $r = 2m$ represents the absolute boundary of the set of all events which can be observed in principle by an external inertial observer. The term « event horizon » is used also in cosmology for essentially the same concept (cf. RINDLER [5]). In the present case the horizon is less observer-dependent than in the cosmological situations, so I shall tend to refer to the hypersurface $r = 2m$ as the ***absolute event horizon*** (**) of the space-time (2).

This, then, is the standard spherically symmetrical collapse picture presented by general relativity. But do we have good reason to trust this picture? Need we believe that it necessarily accords, even in its essentials, with physical reality? Let me consider a number of possible objections:

a) densities in excess of nuclear densities inside,

b) exact vacuum assumed outside,

c) zero net charge and zero magnetic field assumed,

d) rotation excluded,

e) asymmetries excluded,

(*) In fact, if m is of the order of a few solar masses, the tidal forces would already be easily large enough to kill a man in free fall, even at $r = 2m$. But for $m > 10^8 M_\odot$ the tidal effect at $r = 2m$ would be no greater than the tidal effect on a freely falling body near the Earth's surface.

(**) In a general space-time with a well-defined external future infinity, the absolute event horizon would be defined as the boundary of the union of all timelike curves which escape to this external future infinity. In the terminology of PENROSE [6], if $\mathscr{M}$ is a weakly asymptotically simple space-time, for example, then the absolute event horizon in $\mathscr{M}$ is $\dot{I}_{-}[\mathscr{I}^{+}]$.

f) possible λ-term not allowed for,

g) quantum effects not considered,

h) general relativity a largely untested theory,

i) no apparent tie-up with observations.

As regards *a*), it is true that for a body whose mass is of the order of $M_\odot$, its surface would cross $r = 2m$ only after nuclear densities had been somewhat exceeded. It may be argued, then, that too little is understood about the nature of matter at such densities for us to be at all sure how the star would behave while still outside $r = 2m$. But this is not really a significant consideration for our general discussion. It could be of relevance only for the least massive collapsing bodies, if at all. For, the larger the mass involved, the smaller would be the density at which it would be expected to cross $r = 2m$. It could be that very large masses indeed may become involved in gravitational collapse. For $m > 10^{11} M_\odot$ (*e.g.* a good-sized galaxy), the averaged density at which $r = 2m$ is crossed would be less than that of air!

The objections *b*), *c*), *d*), *e*) and, to some extent, *f*) can all be partially handled if we extract, from Fig. 2, only that essential qualitative piece of information which characterizes the solution (2) as describing a collapse which has passed a « point of no return ». I shall consider this in more detail shortly. The upshot will be that if a collapse situation develops in which deviations from (2) near $r = 2m$ at one time are not too great, then two consequences are to be inferred as to the subsequent behaviour. In the first instance an absolute event horizon will arise. Anything which finds itself inside this event horizon will not be able to send signals to the outside worlds. Thus, in this respect at least, the qualitative nature of the « $r = 2m$ » hypersurface in (2) will remain. Similarly, an analogue of the physical singularity at $r = 0$ in (2) will still develop in these more general situations. That is to say, we know from rigorous theorems in general-relativity theory that there must be *some* space-time singularity resulting inside the collapse region. However, we do not know anything about the detailed *nature* of this singularity. There is no reason to believe that it resembles the $r = 0$ singularity of the Schwarzschild solution very closely.

In regard to *c*), *d*) and *f*) we can actually go further in that *exact* solutions are known which generalize the metric (2) to include angular momentum (KERR [7]) and, in addition, charge and magnetic moment (NEWMAN *et al.* [8]), where a cosmological constant may also be incorporated (CARTER [9]). These solutions appear to be somewhat special in that, for example, the gravitational quadrupole moment is fixed in terms of the angular momentum and the mass, while the magnetic-dipole moment is fixed in terms of the angular momentum, charge and mass. However, there are some reasons for believing that these solutions may actually represent the general exterior asymptotic limit resulting

CHAPTER 40

from the type of collapse we are considering. Any extra gravitational multipole moments of quadrupole type, or higher, can be radiated away by gravitational radiation; similarly, extra electromagnetic multipole moments of dipole type, or higher, can be radiated away by electromagnetic radiation. (I shall discuss this a little more later.) If this supposition is correct, then *e*) will to some extent also be covered by an analysis of these exact solutions. Furthermore, *b*) would, in effect, be covered as well, provided we assume that all matter (with the exception of electromagnetic field—if we count that as « matter ») in the neighbourhood of the « black hole », eventually falls into the hole. These exact solutions (for small enough angular momentum, charge and cosmological constant) have absolute event horizons similar to the $r = 2m$ horizon in (2). They also possess space-time curvature singularities, although of a rather different structure from $r = 0$ in (2). However, we would not expect the detailed structure of these singularities to have relevance for a generically perturbed solution in any case.

It should be emphasized that the above discussion is concerned only with collapse situations which do not differ too much initially from the spherically symmetrical case we originally considered. It is not known whether a gravitational collapse of a *qualitatively different character* might not be possible according to general relativity. Also, even if an absolute event horizon *does* arise, there is the question of the « stability » of the horizon. An « unstable » horizon might be envisaged which itself might develop into a curvature singularity. These, again, are questions I shall have to return to later.

As for the possible relevance of gravitational quantum effects, as suggested in *g*), this depends, as far as I can see, on the existence of regions of space-time where there are extraordinary local conditions. If we assume the existence of an absolute event horizon along which curvatures and densities remain small, then it is very hard to believe that a classical discussion of the situation is not amply adequate. It may well be that quantum phenomena have a dominating influence on the physics of the deep interior regions. But whatever effects this might have, they would surely not be observable from the outside. We see from Fig. 2 that such effects would have to propagate outwards in *spacelike* directions over « classical» regions of space-time. However, we must again bear in mind that these remarks might not apply in some qualitatively different type of collapse situation.

We now come to *h*), namely the question of the validity of general relativity in general, and its application to this type of problem in particular. The inadequacy of the observational data has long been a frustration to theorists, but it may be that the situation will change somewhat in the future. There are several very relevant experiments now being performed, or about to be performed. In addition, since it has become increasingly apparent that « strong » gravitational fields probably play an important role in some astrophysical phenomena, there appears to be a whole new potential testing-ground for the theory.

Among the recently performed experiments, designed to test general relativity, one of the most noteworthy has been that of DICKE and GOLDENBERG [10], concerning the solar oblateness. Although the results have seemed to tell against the pure Einstein theory, the interpretations are not really clear-cut and the matter is still somewhat controversial. I do not wish to take sides on this issue. Probably one must wait for further observations before the matter can be settled. However, whatever the final outcome, the oblateness experiment had, for me, the importance of forcing me to examine, once more, the foundations of Einstein's theory, and to ask what parts of the theory are likely to be « here to stay » and what parts are most susceptible to possible modification. Since I feel that the « here to stay » parts include those which were most revolutionary when the theory was first put forward, I feel that it may be worth-while, in a moment, just to run over the reasoning as I see it. The parts of the theory I am referring to are, in fact, the geometrical interpretation of gravity, the curvature of space-time geometry and general-relativistic causality. These, rather than any particular field equations, are the aspects of the theory which give rise to what perhaps appears most immediately strange in the collapse phenomenon. They also provide the physical basis for the major part of the subsequent mathematical discussion.

To begin with, let us agree that it is legitimate to regard space-time as constituting a four-dimensional smooth manifold (or « continuum »). I do not propose to give a justification of this, because on an ordinary macroscopic level it is normally taken as « obvious ». (On the other hand, I think that at a deeper submicroscopic level it is almost certainly « false », but this is not likely to affect the normal discussion of space-time structure—except perhaps *at* a space-time singularity!) Next, we must establish the existence of a *physically well-defined metric* ds which defines for our manifold a (pseudo-) Riemannian structure, with signature ($+ - - -$). The meaning of ds is to be such that when integrated along the world line of any particle, it gives the lapse of proper times as experienced by that particle. Thus, the existence of ds depends on the existence of *accurate clocks* in nature. These clocks must behave locally according to the laws of special relativity. Also, for any two such clocks following the same world line, the time rates they register must agree with one another along the line and should not depend on, say, differing histories for the two clocks. That such clocks do seem to exist in nature, in effect, is a consequence of the fact that any mass m has associated with it a natural frequency mh^{-1}. Thus, the existence of accurate clocks comes down ultimately, via quantum mechanics, to the existence of well-defined masses in nature, whose relative values are in strict proportion throughout space-time. Of course, it might ultimately turn out that the mass ratios of particles are not constant throughout space-time. Then different particles might define slightly different (conformally related) metrics for space-time. But the evidence at present is strongly against any *appreciable* difference existing.

If two neighbouring events in space-time have a separation such that $\mathrm{d}s^2 \geqslant 0$, then according to special relativity, it is possible for one to have a causal influence on the other; if $\mathrm{d}s^2 < 0$, then it is not. We expect this to persist also on a global scale. Thus, it is possible, of two events, for one to influence the other causally if and only if there is a timelike or null curve connecting them.

The existence of a physically well-defined metric and causal structure for space-time, then, seems to be fairly clearly established. It is not so clear, however, that this metric, as so defined, is going to be nonflat. However, we can take the experiment of POUND and REBKA [11] as almost a direct measurement establishing the nonflat nature of space-time. (For this, strictly speaking, the experiment would have to be repeated at various points on the Earth's surface.) The measured $\mathrm{d}s$ near the Earth's surface and the $\mathrm{d}s$ further from the Earth's surface cannot both be incorporated into the same Minkowskian framework because of the « clock slowing » effect (cf. SCHILD [12]). Furthermore, owing to energy balance considerations it is clear that it is with *gravitational* fields that this « clock slowing » effect occurs (owing to the fact that it is *energy, i.e. mass* which responds to a gravitational field). Thus gravitation must be directly related to space-time curvature.

Since we have a (pseudo-) Riemannian manifold, we can use the standard techniques of differential geometry to investigate it. In particular, we can construct a *physically meaningful* Riemann tensor R_{abcd} and thence its Einstein tensor

$$G_{ab} = R_{ab} - \tfrac{1}{2} R g_{ab} \,.$$

Because of the contracted Bianchi identities we know that this satisfies the usual vanishing divergence law. But we also have a symmetric tensor T_{ab}, namely the local energy-momentum tensor (composed of all fields but gravitation), which must satisfy a similar vanishing divergence law. It does not then *necessarily* follow that

$$G_{ab} + \lambda g_{ab} = -8\pi T_{ab} \tag{3}$$

for some constant λ, but it is worth remarking that if we do *not* postulate this equation, then we have not just one, but *two* (linearly unrelated) conserved « energylike » quantities, namely G_{ab} and T_{ab}. In fact, this is just what happens in the theory of BRANS and DICKE [13]. (Such a motivation for the choice of Einstein's field equations (3), does not to my mind have quite the force of the earlier argument, so alternatives to (3) are certainly well worth considering.) Finally, the geodesic motion of monopole test particles may be taken as a consequence of the vanishing divergence condition on T_{ab} (EINSTEIN and GROMMER [14]).

So I want to admit the possibility that Einstein's field equations may be wrong, but *not* (that is, in the macroscopic realm, and where curvatures or densities are not fantastically large) that the general pseudo-Riemannian geometric framework may be wrong. Then the mathematical discussion of the collapse phenomenon can at least be applied. It is interesting that the general mathematical discussion of collapse actually uses very little of the details of Einstein's equations. All that is needed is a certain inequality related to positive-definiteness of energy. In fact, the adoption of the Brans-Dicke theory in place of Einstein's would make virtually no qualitative difference to the collapse discussion.

The final listed objection to the collapse picture is *h*), namely the apparent lack of any tie-up with observed astronomical phenomena. Of course it could be argued that the prediction of the « black hole » picture is simply that we will not see anything—and this is precisely consistent with observations since no « black holes » have been observed! But the real argument is really the other way around. Quasars *are* observed. And they apparently have such large masses and such small sizes that it would seem that gravitational collapse ought to have taken over. But quasars are also long-lived objects. The light they emit does not remotely resemble the exponential cut-off in intensity, with approach to infinite red shift, that might be inferred from the spherically symmetrical discussion. This has led a number of astrophysicists to question the validity of Einstein's theory, at least in its applicability to these situations.

My personal view is that while it is certainly possible (as I have mentioned earlier) that Einstein's equations may be wrong, I feel it would be very premature indeed to dismiss these equations just on the basis of the quasar observations. For, the *theoretical* analysis of collapse, according to Einstein's theory, is still more or less in its infancy. We just do not know, with much certainty, what the consequences of the theory really are. It would be a mistake to fasten attention just on those aspects of general-relativistic collapse which *are* known and to assume that this gives us essentially the complete picture. (It is perhaps noteworthy that many general-relativity theorists have a tendency, themselves, to be a bit on the sceptical side as regards the « classical » collapse picture!) Since it seems to me that there are a number of intriguing largely unexplored possibilities, I feel it may be worth-while to present the « generic » general-relativistic collapse picture as I see it, not only as regards the known theorems, but also in relation to some of the more speculative and conjectural aspects of the situation.

To begin with, let us consider what the general theorems *do* tell us. In order to characterize the situation of collapse « past a point of no return », I shall first need the concept of a *trapped surface.* Let us return to Fig. 2. We ask what qualitative peculiarity of the region $r < 2m$ (after the star has collapsed through) is present. Can such peculiarities be related to the fact that everything appears to be forced inwards in the direction of the centre? It should be stressed again that *apart* from $r = 0$, the space-time at any individual point inside $r = 2m$ is perfectly regular, being as « locally Minkowskian » as any other point (outside $r = 0$). So the peculiarities of the $0 < r < 2m$ region must be of a partially « global » nature. Now consider any point T in the (v, r)-plane of Fig. 2 $(r < 2m)$. Such a point actually represents a *spherical 2-surface* in space-time, this being traced out as the θ, φ co-ordinates vary. The surface area of this sphere is $4\pi r^2$. We imagine a flash of light emitted simultaneously over this spherical surface T. For an ordinary spacelike 2-sphere in flat space-time, this would result in an ingoing flash imploding towards the centre (surface area decreasing) together with an outgoing flash exploding outwards (surface area increasing). However, with the surface T, while we still have an ingoing flash with decreasing surface area as before, the « outgoing » flash, on the other hand, is in effect also falling inwards (though not as rapidly) and its surface area also decreases. The surface T ($v =$ const, $r =$ const $< 2m$) of metric (2) serves as the prototype of a trapped surface. If we perturb the metric (2) slightly, in the neighbourhood of an initial hypersurface, then we would still expect to get a surface T with the following property:

> T is a spacelike closed (*) 2-surface such that the null geodesics which meet it orthogonally all *converge* initially at T.

This convergence is taken in the sense that the local surface area of cross-section *decreases*, in the neighbourhood of each point of T, as we proceed into the future. (These null geodesics generate, near T, the boundary of the set of points lying causally to the future of the set T.) Such a T is called a *trapped surface.*

We may ask whether any connection is to be expected between the existence of a trapped surface and the presence of a physical space-time singularity such as that occurring at $r = 0$ in (2). The answer supplied by some general theorems (PENROSE [6, 15], HAWKING and PENROSE [16]) is, in effect, that the presence of a trapped surface always *does* imply the presence of some form of space-time singularity.

There are similar theorems that can also be applied in cosmological situations. For example (HAWKING [17-19], HAWKING and PENROSE [16]), if the universe is *spatially closed,* then (excluding exceptional limiting cases, and assuming $\lambda \leqslant 0$) the conclusion is that there must be a space-time singularity. This time we expect the singularity to reside in the past (the « big bang »). Other theorems (HAWKING [17, 19], HAWKING and PENROSE [16]) can be applied also to spatially open universes. For example, if there is any point

---

(*) By a « closed » surface, hypersurface, or curve, I mean one that is « compact without boundary ».

(*e.g.* the Earth at the present epoch) whose past light cone starts « converging again » somewhere in the past (*i.e.* objects of given size start to have *larger* apparent angular diameters again when their distance from us exceeds some critical value), then, as before, the presence of space-time singularities is implied ($\lambda \leqslant 0$). According to HAWKING and ELLIS ([20], cf. also HAWKING and PENROSE [16]) the presence and isotropy of the 3°K radiation strongly indicates that the above condition on our past light cone is actually satisfied. So the problem of space-time singularities does seem to be very relevant to our universe, also on a large scale.

The main significance of theorems such as the above, is that they show that the presence of space-time singularities in exact models is not just a feature of their high symmetry, but can be expected also in generically perturbed models. This is *not* to say that *all* general-relativistic curved space-times are singular—far from it. There are many exact models known which are complete and free from singularity. But those which resemble the standard Friedmann models or the Schwarzschild collapse model sufficiently closely must be expected to be singular ($\lambda \leqslant 0$). The hope had often been expressed (cf. LINDQUIST and WHEELER [21], LIFSHITZ and KHALATNIKOV [22]) that the actual space-time singularity occurring in a collapsing space-time model might have been a consequence more of the fact that the matter was all hurtling simultaneously towards one central point, than of some intrinsic feature of general-relativistic space-time models. When perturbations are introduced into the collapse, so the argument could go, the particles coming from different directions might « miss » each other, so that an effective « bounce » might ensue. Thus, for example, one might envisage an « oscillating » universe which on a *large* scale resembles the cycloidal singular behaviour of an « oscillating » spatially closed Friedmann model; but the *detailed* behaviour, although perhaps involving enormous densities while at maximal contraction, might, by virtue of complicated asymmetries, contrive to avoid actual space-time singularities. However, the theorems seem to have ruled out a singularity-free « bounce » of this kind. But the theorems do *not* say that the singularities need resemble those of the Friedmann or Schwarzschild solutions at all closely. There is some evidence (cf. MISNER [23], for example) that the « generic » singularities may be very elaborate and possess a qualitative structure very different from that of their smoothed-out counterparts. Very little is known about this, however.

It is worth mentioning the essential basic assumptions that enter into the theorems. In the first place we require an « energy condition » which, by virtue of Einstein's equations (3), may be stated as a negative-definiteness condition on the Ricci tensor:

$$t^a t_a = 1 \quad \text{implies} \quad R_{ab} t^a t^b \leqslant 0, \tag{4}$$

that is to say, the time-time component R_{00} of R_{ab} is nonpositive in any orthonormal frame. If we assume $\lambda = 0$ in Einstein's equations (3), then (4) becomes

$$t^a t_a = 1 \quad \text{implies} \quad T_{ab} t^a t^b \geqslant \tfrac{1}{2} T^c_c \,.$$

This, when referred to an eigenframe of T_{ab}, can be stated as

$$E + p_\alpha \geqslant 0 \quad \text{and} \quad E + \sum p_\alpha \geqslant 0 \,, \tag{5}$$

where $\alpha = 1, 2, 3$. Here E is the energy density (referred to this frame) and p_1, p_2, p_3 are the three principal pressures. If (3) holds with $\lambda < 0$, then it is still true that (4) is a *consequence* of (5). The significance of the energy condition (4) lies in the effect of RAYCHAUDHURI [24] which states that whenever a system of timelike geodesics normal to a spacelike hypersurface starts converging, then this convergence inevitably increases along the geodesics until finally the geodesics cross over one another (assuming the geodesics are complete).

There is a corresponding focussing effect in the case of *null* geodesics. This depends on the « weak energy condition »:

$$l^a l_a = 0 \quad \text{implies} \quad R_{ab} l^a l^b \leqslant 0 \,. \tag{6}$$

This condition (6) is a consequence of (4) (as follows by a limiting argument) but not conversely. If we assume Einstein's equations (3) with, now, *any value of* λ, then (6) is equivalent to

$$E + p_\alpha \geqslant 0 \tag{7}$$

for $\alpha = 1, 2, 3$. The conditions (7) are, in fact, a consequence of the *nonnegative definiteness of the energy density*:

$$t^a t_a = 1 \quad \text{implies} \quad T_{ab} t^a t^b \geqslant 0$$

(that is $T_{00} \geqslant 0$ in each orthonormal frame). Thus, there is a strong physical basis for (6). The physical basis for (4) is not quite so strong, but provided $\lambda \leqslant 0$, we would certainly expect (4) to hold for all normal matter. (Note that if $E > 0 \geqslant \lambda$, only large *negative* pressures could cause trouble with (4). Usually people only worry about large positive pressures!) It is the « strong » condition (4) that is required for the proofs of most of the theorems, but much can be said, concerning the qualitative nature of a collapse situation, even on the basis of the « weak » condition (6) alone (cf. PENROSE [15]).

A remark concerning the condition on the cosmological constant λ seems appropriate here. It is a weakness of the theorems that most of them do require $\lambda \leqslant 0$ for their strict applicability. However, it would appear that the condition $\lambda \leqslant 0$ is only really relevant to the initial setting of the global conditions on the space-time which are required for applicability of the theorem. If curvatures are to become large near a singularity, then (from dimensional considerations alone) the λ-term will become more and more insignificant. So it seems unlikely that a λ-term will really make much difference to the singularity structure in a collapse. The relevance of λ is really only at the cosmological scale.

Most of the theorems (but not all, cf. HAWKING [19]) require, as an additional assumption, the nonexistence of closed timelike curves. This is a very reasonable requirement, since a space-time which possesses closed timelike curves would allow an observer to travel into his own past. This would lead to very serious interpretative difficulties! Even if it could be argued, say, that the accelerations involved might be such as to make the trip impossible in « practice » (cf. GÖDEL [25]), equally serious difficulties would arise for the observer if he merely reflected some light signals into his own past! In addition closed timelike curves can lead to unreasonable consistency conditions on the solutions of hyperbolic differential equations. In any case, it seems unlikely that closed timelike curves can substitute for a space-time singularity, except in special unstable models.

Some of the theorems require an additional « generality » condition, to the effect that every timelike or null geodesic enters some region in which the curvature is not everywhere lined up in a particular way with the geodesic. (More precisely, $t_{[a}R_{b]cd[e}t_{f]}t^c t^d \neq 0$ somewhere along the geodesic, t^a being its tangent vector.) This condition plays a role in the mathematics, but from the physical point of view it is really no condition at all. We would always expect a little bit of matter or randomly oriented curvature along any geodesic in a physically realistic solution. It is only in very special limiting cases that we would expect the condition to be violated. (Curiously enough, however, practically every explicitly known solution does violate the condition!)

Finally, it should be remarked that none of the theorems *directly* establishes the existence of regions of approaching infinite curvature. Instead, all one obtains is that the space-time is not geodesically complete (in timelike or null directions) and, furthermore, cannot be *extended* to a geodesically complete space-time. (« Geodesically complete » means that geodesics can be extended indefinitely to arbitrarily large values of their length or affine parameter—so that inertially moving particles or photons do not just « fall off the edge » of the space-time.) The most « reasonable » explanation for why the space-time is not inextendible to a complete space-time seems to be (and I would myself believe this to be the most likely, in general) that the space-time is confronted with, in some sense, *infinite curvature* at its boundary. But the theorems do not quite say this. Other types of space-time singularity are possible, and theorems of a somewhat different nature would be required to decide which is the most likely type of singularity to occur.

We must now ask the question whether the theorems are actually likely to be relevant in the case of a collapsing star or superstar. Do we, in fact, have any reason to believe that trapped surfaces can ever arise in gravitational collapse? I think a very strong case can be made that at least *sometimes* a trapped surface must arise. I would not expect trapped surfaces necessarily always to arise in a collapse. It might depend on the details of the situation. But if we can establish that there can be nothing *in principle* against a trapped surface arising—even if in some very contrived and outlandish situation—then we must surely accept that trapped surfaces must at least occasionally arise in real collapse.

Rather than use the trapped-surface condition, however, it will actually be somewhat easier to use the alternative condition of the existence of a point whose light cone starts « converging again ». From the point of view of the general theorems, it really makes no essential difference which of the two conditions is used. Space-time singularities are to be expected in either case. Since we are here interested in a collapse situation rather than in the « big bang », we shall be concerned with the *future* light cone C of some point p. What we have to show is that it is possible in principle for enough matter to cross to within C, so that the divergence of the null geodesics which generate C changes sign somewhere to the future of p. Once these null geodesics start to converge, then « weak energy condition » (6) will take over, with the implication that an absolute event horizon must develop (outside C). As a consequence of the stronger « energy condition » (4) it will also follow that space-time singularities will occur.

Since we ask only that it be possible *in principle* to reconverge the null rays generating C, we can resort to an (admittedly far-fetched) « gedanken experiment ». Consider an elliptical galaxy containing, say, 10^{11} stars. Suppose, then, that we contrive to alter the motion of the stars slightly by eliminating the transverse component of their velocities. The stars will then fall inwards towards the centre. We may arrange to steer them, if we like, so as to ensure that they all reach the vicinity of the centre at about the same time without colliding with other stars. We only need to get them into a volume of diameter about fifty times that of the solar system, which gives us plenty of room for all the stars. The point p is now taken near the centre at about the time the stars enter this volume (Fig. 3). It is easily seen from the orders of magnitude involved, that the relativistic light deflection (an *observed* effect of general relativity) will be sufficient to cause the null rays in C to reconverge, thus achieving our purpose.

Let us take it, then, that absolute event horizons can sometimes occur in a gravitational collapse. Can we say anything more detailed about the nature of the resulting situation? Hopeless as this problem may appear at first sight, I think there is actually a reasonable chance that it may find a large measure of solution in the not-too-distant future. This would depend on the validity of a certain result which has been independently conjectured by a number of people. I shall refer to this as the *generalized* (*) *Israel conjecture* (abbreviated GIC). Essentially GIC would state: if an absolute event horizon develops in an asymptotically flat space-time, then the solution exterior to this horizon approaches a Kerr-Newman solution asymptotically with time.

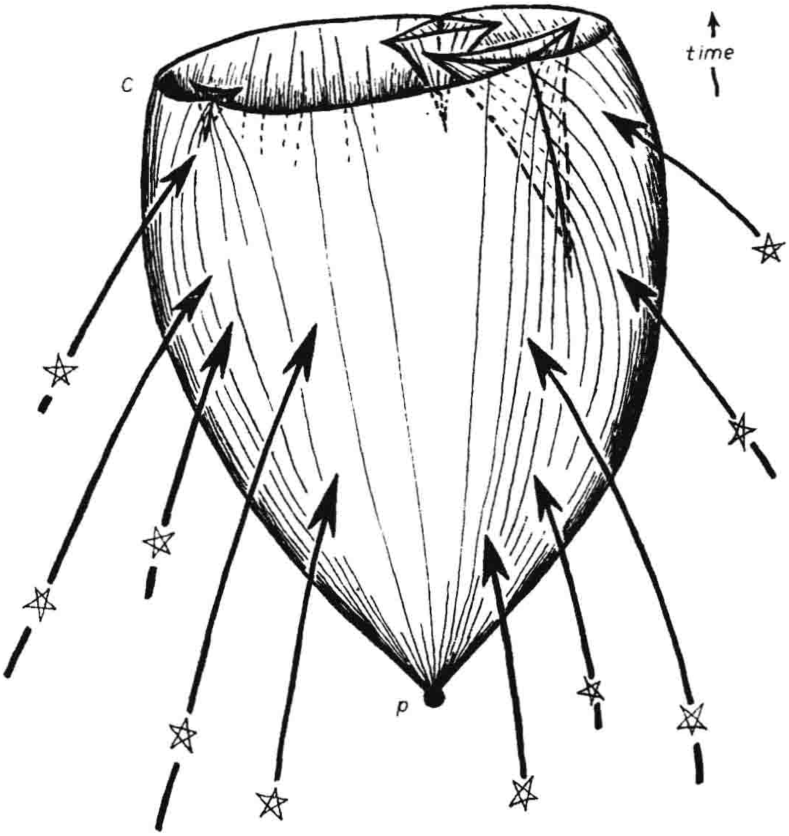


Fig. 3. – The future light cone of p is caused to reconverge by the falling stars.

The Kerr-Newman solutions (KERR [7], NEWMAN *et al.* [8]) are explicit asymptotically flat stationary solutions of the Einstein-Maxwell equation ($\lambda = 0$) involving just *three* free parameters m, a and e. As with the metric (1), the *mass*, as measured asymptotically, is the parameter m (in gravitational units). The solution also possesses angular momentum, of magnitude am. Finally, the total charge is given by e. When $a = e = 0$ we get the Schwarzschild solu-

---

(*) ISRAEL conjectured this result only in the stationary case, hence the qualification « generalized ». In fact, ISRAEL has expressed sentiments opposed to GIC. However, Israel's theorem [26, 27] represents an important step towards establishing of GIC, if the conjecture turns out to be true.

tion. Provided that

$$m^2 \geqslant a^2 + e^2$$

the solution has an absolute event horizon. CARTER [9] has shown how to obtain all the geodesics and charged orbits for this solution, reducing the problem to a single quadrature. Thus, if GIC is true, then we shall have remarkably complete information as to the asymptotic state of affairs resulting from a gravitational collapse.

But what reason is there for believing that GIC has any chance of being true? One indication comes from a perturbation analysis of the Schwarzschild solution (REGGE and WHEELER [28], DOROSHKEVICH *et al.* [29]) which seems to indicate that all perturbations except rotation have a tendency to be damped out. Another indication is the theorem of ISRAEL [26] which states, in effect, that the Schwarzschild solution is the only static asymptotically flat vacuum solution with an absolute event horizon (although there is a nontrivial side-condition to the theorem; cf. also THORNE [30] for the axially symmetric case). ISRAEL [27] has also generalized his result to the Einstein-Maxwell theory, finding the spherically symmetric Reissner-Nordstrom solution to be the only asymptotically flat static solution with an absolute event horizon. CARTER [31] has made some progress, in the vacuum rotating case, towards the objective of establishing the Kerr solution ($e = 0$) as the general asymptotically flat stationary solution with an absolute event horizon. In addition, there are solutions of the vacuum equations known (ROBINSON and TRAUTMAN [32]), which are suitably asymptotically flat and nonrotating, which apparently possess absolute event horizons, but are nonstatic. As time progresses they become more and more symmetrical, approaching the Schwarzschild solution asymptotically with time [33]. In the process, the higher multipole moments are radiated away as gravitational radiation.

The following picture then suggests itself. A body, or collection of bodies, collapses down to a size comparable to its Schwarzschild radius, after which a trapped surface can be found in the region surrounding the matter. Some way outside the trapped surface region is a surface which will ultimately be the absolute event horizon. But at present, this surface is still expanding somewhat. Its exact location is a complicated affair and it depends on how much more matter (or radiation) ultimately falls in. We assume only a finite amount falls in and that GIC is true. Then the expansion of the absolute event horizon gradually slows down to stationarity. Ultimately the field settles down to becoming a Kerr solution (in the vacuum case) or a Kerr-Newman solution (if a nonzero net charge is trapped in the « black hole »).

Doubts have frequently been expressed concerning GIC, since it is felt that a body would be unlikely to throw off all its excess multipole moments just as it crosses the Schwarzschild radius. But with the picture presented above this is not necessary. I would certainly not expect the body itself to throw off its multipole moments. On the other hand, the gravitational field *itself* has a lot of settling-down to do after the body has fallen into the « hole ». The asymptotic measurement of the multipole moments need have very little to do with the detailed structure of the body itself; the *field* can contribute very significantly. In the process of settling down, the field radiates gravitationally—and electromagnetically too, if electromagnetic field is present. Only the mass, angular momentum and charge need survive as ultimate independent parameters. (Presumably the charge parameter e would be likely to be very small by comparison with a and m.)

But suppose GIC is not true, what then? Of course, it may be that there are just a lot more possible limiting solutions than that of Kerr-Newman. This would mean that much more work would have to be done to obtain the detailed picture, but it would not imply any qualitative change in the set-up. On the other hand there is the more alarming possibility that the absolute event horizon may be *unstable*! By this I mean that instead of settling down to become a nice smooth solution, the space-time might gradually develop larger and larger curvatures in the neighbourhood of the absolute event horizon, ultimately to become effectively singular there. My personal opinion is that GIC is more likely than this, but various authors have expressed the contrary view (*).

If such instabilities are present then this would certainly have astrophysical implications. But even if GIC is true, the resulting « black hole » may by no means be so « dead » as has often been suggested. Let us examine the Kerr-Newman solutions, in the case $m^2 > a^2 + e^2$ in a little more detail. But before doing so let us refer back to the Schwarzschild solution (2). In Fig. 4, I have drawn what is, in effect, a cross-section of the space-time, given by $v - r =$ const. The circles represent the location of a flash of light which had been emitted at the nearby point a moment earlier. Thus, they indicate the orientation of the light cones in the space-time. We note that for large r the point lies inside the circle, which is consistent with the static nature of the space-time (*i.e.* one can « stay in the same place » while retaining a timelike world line). On the other hand for $r < 2m$ the point lies outside the circle, indicating that all matter must be dragged inwards if it is to remain moving in a timelike direction (so, to « stay in the same place » one would have to exceed the local speed of light). Let us now consider the corresponding picture for the Kerr-Newman solutions with $m^2 > a^2 + e^2$ (Fig. 5). I shall not be concerned, here, with the curious nature of the solution inside the absolute

---

(*) Some recent work of NEWMAN [34] on the charged Robinson-Trautman solutions suggests that new features indicating instabilities may arise when an electromagnetic field is present.

event horizon H, since this may not be relevant to GIC. The horizon H itself is represented as a surface which is tangential to the light cones at each of its points. Some distance outside H is the « stationary limit » L, at which

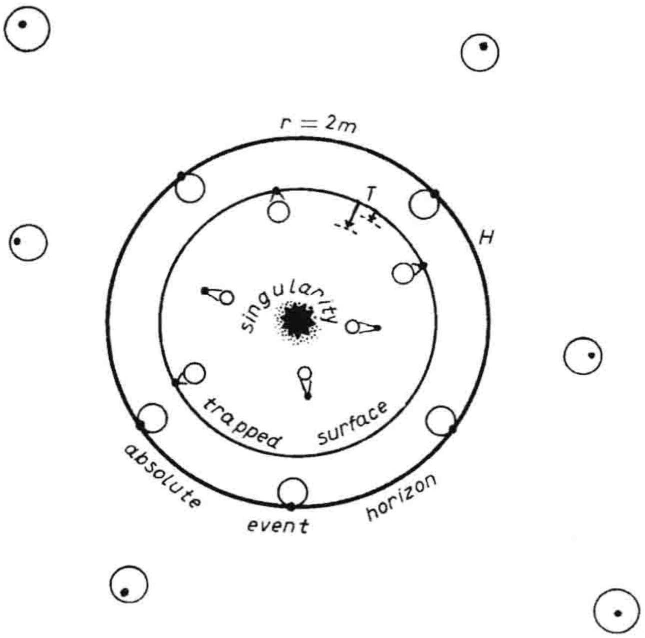


Fig. 4. – Spatial view of spherical « black hole » (Schwarzschild solution).

one must travel with the local light velocity in order to « stay in the same place ».

I want to consider the question of whether it is possible to extract energy out of a « black hole ». One might imagine that, since the matter which has fallen through has been lost for ever, so also is its energy content irrevocably trapped. However, it is not totally clear to me that this need be the case. There are at least two methods (neither of which is very practical) which might be construed as mechanisms for extracting energy from a « black hole ». The first is due to MISNER [35]. This requires, in fact, a whole *galaxy* of 2^N « black holes », each of mass m. We first bring them together in pairs and allow them to spiral around one another, ultimately to swallow each other up. During the spiraling, a certain fraction K of their mass-energy content is radiated away as gravitational energy, so the mass of the resulting « black hole » is $2m(1-K)$. The energy of the gravitational waves is collected and the process is repeated. Owing to the scale invariance of the gravitational vacuum equations, the same fraction of the mass-energy is collected in the form of gravitational waves at each stage. Finally we end up with a single « black hole » of mass $2^N m(1-K)^N$. Now, the point is that *however small* K may in fact be, we can always choose N large enough so that $(1-K)^N$ is as small as we please. Thus, in principle, we can extract an arbitrarily large fraction of the mass-energy content of Misner's galaxy.

But anyone at all familiar with the problems of detecting gravitational

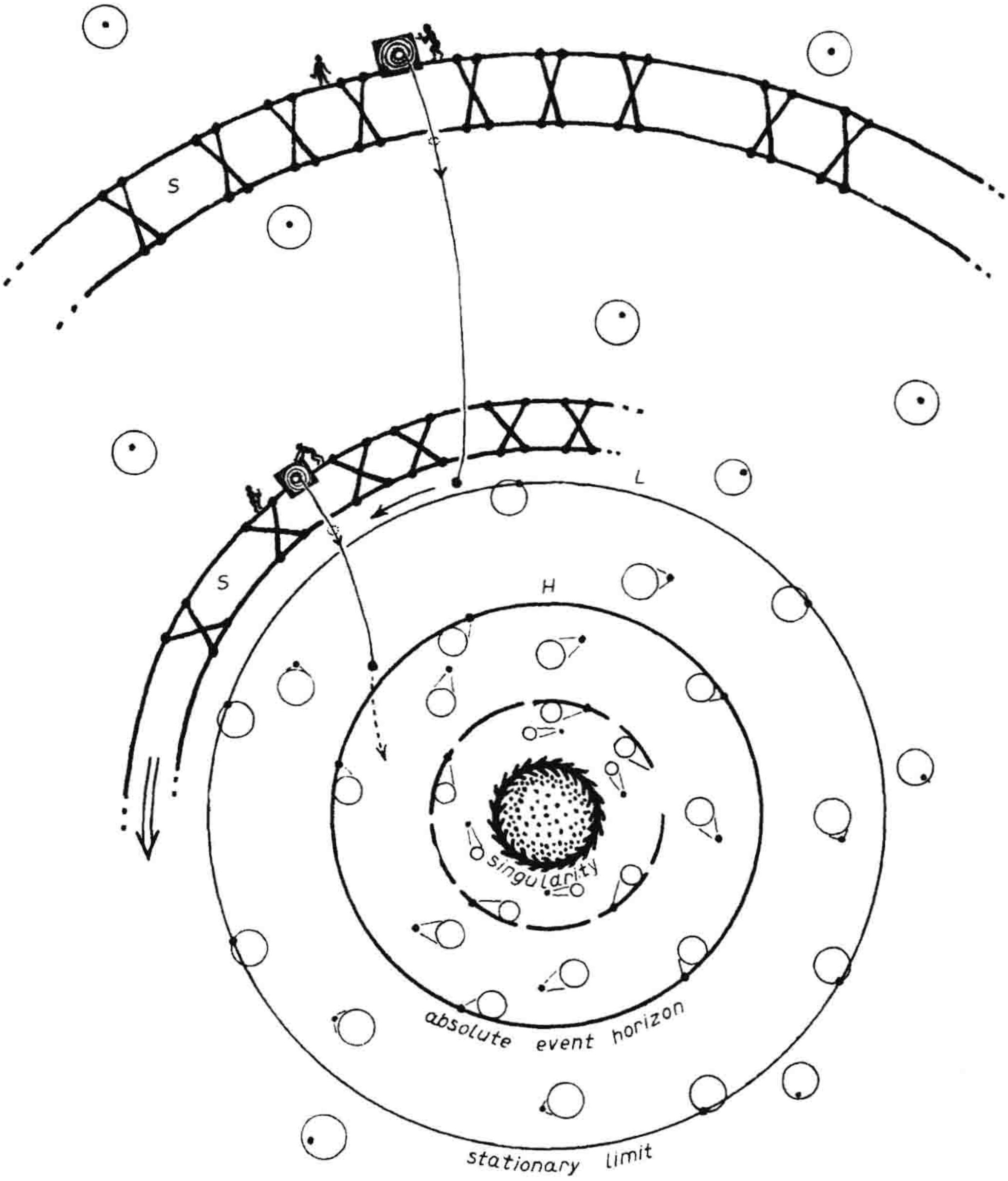


Fig. 5. – Rotating « black hole » (Kerr-Newman solution with $m^2 > a^2 + e^2$). The inhabitants of the structures S and S^* are extracting rotational energy from the « black hole ».

radiation will be aware of certain difficulties! Let me suggest another method which actually tries to do something a little different, namely extract the « rotational energy » of a « rotating black hole » (Kerr solution). Consider Fig. 5 again. We imagine a civilization which has built some form of stabilized structure S surrounding the « black hole ». If they lower a mass slowly on a (light, inextensible, unbreakable) rope until it reaches L, they will be able to recover, at S, the entire energy content of the mass. If the mass is released as it reaches L then they will simply have bartered the mass for its energy content. (This is the highest-grade energy, however, namely wound-up springs!) But they can do better than this! They also build another structure S^*, which rotates, to some extent, with the « black hole ». The lowering process is continued, using S^*, to beyond L. Finally the mass is dropped through H, but in such a way that its energy content, as measured from S, is *negative*! Thus, the inhabitants of S are able, in effect, to lower masses into the « black hole » in such a way that they obtain *more* than the energy content of the mass. Thus they extract some of the energy content of the « black hole » itself in the process. If we examine this in detail, however, we find that the angular momentum of the « black hole » is also reduced.

Thus, in a sense, we have found a way of extracting *rotational energy* from the « black hole ». Of course, this is hardly a practical method! Certain improvements may be possible, *e.g.*, using a ballistic method (*). But the real significance is to find out what can and what cannot be done *in principle* since this may have some indirect relevance to astrophysical situations.

Let me conclude by making a few highly speculative remarks. In the first place, suppose we take what might be referred to, now, as the most « conservative » point of view available to us, namely that GIC is not only true, but it also represents the *only* type of situation that can result from a gravitational collapse. Does it follow, then, that nothing of very great astrophysical interest is likely to arise out of collapse? Do we merely deduce the existence of a few additional dark « objects » which do little else but contribute, slightly, to the overall mass density of the universe? Or might it be that such « objects », while themselves hidden from direct observation, could play some sort of catalytic role in producing observable effects on a much larger scale. The « seeding » of galaxies is one possibility which springs to mind. And if « black holes » are born of violent events, might they not occasionally be ejected with high velocities when such events occur! (The one thing we can be sure about is that they *would* hold together!) I do not really want to make any very specific

---

(*) Calculations show that this can indeed be done. A particle p_0 is thrown from S into the region between L and H, at which point the particle splits into two particles p_1 and p_2. The particle p_2 crosses H, but p_1 escapes back to S possessing *more* mass-energy content than p_0!

suggestions here. I only wish to make a plea for « black holes » to be taken seriously and their consequences to be explored in full detail. For who is to say, without careful study, that they cannot play some important part in the shaping of observed phenomena?

But need we be so cautious as this? Even if GIC, or something like it, is true, have we any right to suggest that the *only* type of collapse which can occur is one in which the space-time singularities lie hidden, deep inside the protective shielding of an absolute event horizon? In this connection it is worth examining the Kerr-Newman solutions for which $m^2 < a^2 + e^2$. **The situation is depicted in Fig. 6. The absolute event horizon has now completely disappeared! A region of space-time singularity still exists in the vicinity of the centre, but now it is possible for information to escape from the singularity**

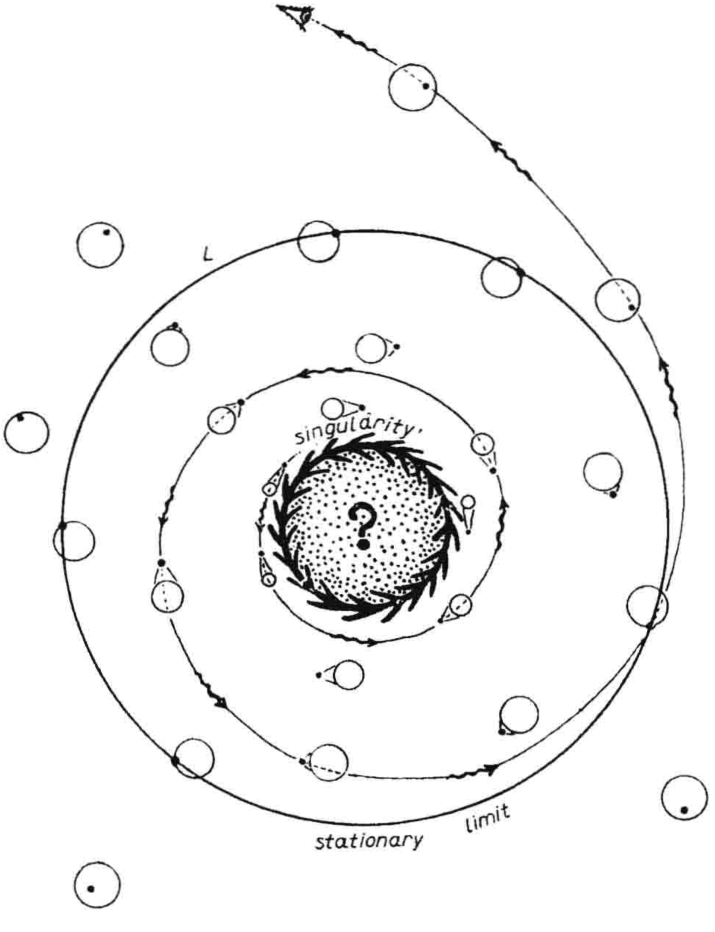


Fig. 6. – A « naked singularity » (Kerr-Newman solution with $m^2 < a^2 + e^2$).

to the outside world, provided it spirals around sufficiently. In short, the singularity is *visibile*, in all its nakedness, to the outside world!

However, there is an essential difference between the logical status of the singularity marked at the centre of Fig. 6 and that marked at the centres of Fig. 4 and 5. In the cases of Fig. 4 and 5 there are trapped surfaces present, so we have a *theorem* which tells us that even with generic perturbation a singularity will still exist. In the situation of Fig. 6, however, we have no trapped surfaces, no known theorem guaranteeing singularities and certainly no analogue of GIC. So it is really an open question whether a situation remotely resembling Fig. 6 is ever likely to arise.

We are thus presented with what is perhaps the most fundamental unanswered question of general-relativistic collapse theory, namely: does there exist a « cosmic censor » who forbids the appearance of naked singularities, clothing each one in an absolute event horizon? In one sense, a « cosmic censor » can be shown *not* to exist. For it follows from a theorem of HAWKING [19] that the « big bang » singularity is, in principle, observable. But it is not known whether singularities observable from outside will ever arise in a generic *collapse*, which starts off from a perfectly reasonable nonsingular initial state.

If in fact naked singularities do arise, then there is a whole new realm opened up for wild speculations! Let me just make a few remarks. If we envisage an isolated naked singularity as a source of new matter in the universe, then we do not *quite* have unlimited freedom in this! For although in the neighbourhood of the singularity we have no equations, we still have normal physics holding in the space-time *surrounding* the singularity. From the mass-energy flux theorem of BONDI *et al.* [36] and SACHS [37], it follows that it is *not* possible for *more* mass to be ejected from a singularity than the original total mass of the system, *unless* we are allowed to be left with a singularity of *negative* total mass. (Such a singularity would *repel* all other bodies, but would still be attracted by them!)

While in the realm of speculation concerning matter production at singularities, perhaps one further speculative remark would not be entirely out of place. This is with respect to the manifest large-scale time asymmetry in the behaviour of matter in the universe (and also the apparent large-scale asymmetry between matter and antimatter). It is often argued that small observed violations of T (and C) invariance in fundamental interactions can have no bearing on the cosmological asymmetry problem. But it is not at all clear to me that this is necessarily so. It is a space-time singularity (*i.e.* presumably the « big bang ») which appears to govern the production of matter in the universe. When curvatures are fantastically large—as they surely are at a singularity—the local physics will be drastically altered. Can one be sure that the asymmetries of local interactions will not have the effect of being as drastically magnified?

When so little is known about the geometrical nature of space-time singularities and even less about the nature of the physics which takes place there, it is perhaps futile to speculate in this way about them. However, ultimately a theory will have to be found to cope with the situation. The question of the quantization of general relativity is often brought up in this connection. My own feeling is that the purpose of correctly combining quantum theory with general relativity is really somewhat different. It is simply a step in the direction of discovering how nature fits together as a whole. When eventually we have a better theory of nature, then perhaps we can try our hands, again, at understanding the extraordinary physics which must take place at a space-time singularity.

REFERENCES

[1] S. Chandrasekhar: *Roy. Astr. Soc. Month. Not.*, **95**, 207 (1935).
[2] J. R. Oppenheimer and G. Volkoff: *Phys. Rev.*, **55**, 374 (1939).
[3] A. S. Eddington: *Nature*, **113**, 192 (1924).
[4] D. Finkelstein: *Phys. Rev.*, **110**, 965 (1958).
[5] W. Rindler: *Roy. Astr. Soc. Month. Not.*, **116**, 6 (1956).
[6] R. Penrose: in *Battelle Rencontres* (ed. C. M. De Witt and J. A. Wheeler) (New York, 1968).
[7] R. P. Kerr: *Phys. Rev. Lett.*, **11**, 237 (1963).
[8] E. T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash and R. Torrence: *Journ. Math. Phys.*, **6**, 918 (1965).
[9] B. Carter: *Phys. Rev.*, **174**, 1559 (1968).
[10] R. H. Dicke and H. M. Goldenberg: *Phys. Rev. Lett.*, **18**, 313 (1967).
[11] R. V. Pound and G. A. Rebka: *Phys. Rev. Lett.*, **4**, 337 (1960).
[12] A. Schild: in *Relativity Theory and Astrophysics*, Vol. **1**: *Relativity and Cosmology* (ed. J. Ehlers) (Providence, R. I., 1967).
[13] C. Brans and R. H. Dicke: *Phys. Rev.*, **124**, 925 (1961).
[14] A. Einstein and J. Grommer: *S. B. Preuss. Akad. Wiss.*, **1**, 2 (1927).
[15] R. Penrose: *Phys. Rev. Lett.*, **14**, 57 (1965).
[16] S. W. Hawking and R. Penrose: *Proc. Roy. Soc.*, A (in press) (1969).
[17] S. W. Hawking: *Proc. Roy. Soc.*, A **294**, 511 (1966).
[18] S. W. Hawking: *Proc. Roy. Soc.*, A **295**, 490 (1966).
[19] S. W. Hawking: *Proc. Roy. Soc.*, A **300**, 187 (1967).
[20] S. W. Hawking and G. F. R. Ellis: *Astrophys. Journ.*, **152**, 25 (1968).
[21] R. W. Lindquist and J. A. Wheeler: *Rev. Mod. Phys.*, **29**, 432 (1957).
[22] E. M. Lifshitz and I. M. Khalatnikov: *Advances in Phys.*, **12**, 185 (1963).
[23] C. W. Misner: *Phys. Rev. Lett.*, **22**, 1071 (1969).
[24] A. K. Raychaudhuri: *Phys. Rev.*, **98**, 1123 (1955).
[25] K. Gödel: in *Albert Einstein Philosopher Scientist*, edited by P. A. Schlipp (New York, 1959), p. 557.
[26] W. Israel: *Phys. Rev.*, **164**, 1776 (1967).
[27] W. Israel: *Commun. Math. Phys.*, **8**, 245 (1968).

[28] T. REGGE and J. A. WHEELER: *Phys. Rev.*, **108**, 1063 (1957).
[29] A. G. DOROSHKEVICH, YA. B. ZEL'DOVICH and I. D. NOVIKOV: *Žurn. Ėksp. Teor. Fiz.*, **49**, 170 (1965); English trans., *Sov. Phys. JETP*, **22**, 122 (1966).
[30] K. S. THORNE: Ph. D. thesis, Princeton University Princeton, N. J. (1965).
[31] B. CARTER: personal communication (1969).
[32] I. ROBINSON and A. TRAUTMAN: *Proc. Roy. Soc.*, A **265**, 463 (1962).
[33] J. FOSTER and E. T. NEWMAN: *Journ. Math. Phys.*, **8**, 189 (1967).
[34] E. T. NEWMAN: personal communication (1969).
[35] C. W. MISNER: personal communication (1968).
[36] H. BONDI, M. G. J. VAN DER BURG and A. W. K. METZNER: *Proc. Roy. Soc.*, A **269**, 21 (1962).
[37] R. K. SACHS: *Proc. Roy. Soc.*, A **270**, 103 (1962).

41

CONSERVED QUANTITIES IN THE EINSTEIN–MAXWELL THEORY

J. Math. Phys. **10(9)** (1969), 1566–70. With A. R. Exton and E. T. Newman

This generalizes the results of Chapters 23 and 35, showing how the NP quantities of those quantities can be modified so that they remain constant in the presence of electromagnetism.

Conserved Quantities in the Einstein–Maxwell Theory*

ALBERT R. EXTON† AND EZRA T. NEWMAN
University of Pittsburgh, Pittsburgh, Pennsylvania
AND
ROGER PENROSE
Birkbeck College,‡ University of London, London, England

(Received 8 March 1967)

It is shown that the 10 gravitationally-conserved quantities defined in asymptotically flat, empty, space-times are, when suitably modified, also conserved in asymptotically flat Einstein–Maxwell space-times. Furthermore, the implied selection rules for transitions between stationary Einstein–Maxwell states are the same as those in the pure gravitational case.

I. INTRODUCTION

It has recently been shown[1,2] that in Einstein vacuum gravitational theory there are ten quantities, defined for asymptotically flat space–times, which are absolutely conserved. In the Einstein–Maxwell theory there are six conserved quantities which are the analogs of the above for the Maxwell field. In a stationary vacuum field, these ten conserved quantities can be expressed as certain combinations of the mass, dipole, and quadrupole moments. This fact leads to a selection rule for purely gravitational transitions between stationary states, namely, that if a field is initially stationary and then becomes radiative, if it is ever to become stationary again, the combination of moments mentioned above must return to its original value.

The purpose of this paper is to show that in asymptotically flat Einstein–Maxwell space–times there are 16 conserved quantities—the six mentioned in Refs. 1 and 2 and, in addition, ten which are closely related to the ten conserved quantities in the pure gravitational case. In stationary states, these ten quantities reduce to the same combination of moments as in the vacuum case. Thus, we have the same selection rules for transitions between stationary Einstein–Maxwell states as for vacuum states with additional selection rules arising from the six Maxwell conserved quantities.

II. ASYMPTOTIC BEHAVIOR

The conservation laws reported in Refs. 1 and 2 were arrived at by consideration of the asymptotic behavior of the fields in question. We shall, therefore, investigate the asymptotic behavior of the Einstein–Maxwell field using the method of Newman–Unti (N–U).[3] We adopt essentially[4] the same coordinate and tetrad conditions used there, where, of course, we use the field equations and Bianchi Identities appropriate to the Einstein–Maxwell field as given by Newman and Penrose.[5] Using the results of Kozarzewski[6] for the asymptotic behavior of the spin coefficients and the peeling-off property of the fields, one obtains by the method of N–U the following asymptotic form for the Einstein–Maxwell field. If

$$\psi_0 = \psi_0^0 r^{-5} + \psi_0^1 r^{-6} + O(r^{-7}), \tag{1}$$

$$\phi_0 = \phi_0^0 r^{-3} + \phi_0^1 r^{-4} + O(r^{-5}), \tag{2}$$

and the various u, θ, φ derivatives of these expressions also hold, then

$$\psi_1 = \psi_1^0 r^{-4} + (\bar{\eth}\psi_0^0 + 3\bar{\phi}_1^0\phi_0^0)r^{-5} + O(r^{-6}), \tag{3}$$

$$\psi_2 = \psi_2^0 r^{-3} + (\bar{\eth}\psi_1^0 + 2\phi_1^0\bar{\phi}_1^0)r^{-4} + O(r^{-5}), \tag{4}$$

---

* Work supported, in part, by the Aerospace Research Laboratories, Office of Aerospace Research, U.S. Air Force.

† Andrew Mellon Fellow. Present address: The University of Texas, Austin, Texas.

‡ This work was done while the author was at Yeshiva and Princeton Universities.

[1] E. Newman and R. Penrose, Phys. Rev. Letters 15, 231 (1965).

[2] E. Newman and R. Penrose, Proc. Roy. Soc. (London) 305, 175 (1968).

[3] E. Newman and T. Unti, J. Math. Phys. 3, 891 (1962).

[4] Some slight changes in the notation from that of Refs. 3 and 8 are being introduced here (and in Ref. 2), although agreement with Ref. 7 is being maintained. This results in a somewhat simplified appearance of most of the equations; in particular, the removal of all factors of $(2)^{\frac{1}{2}}$. To emphasize that the notation has been altered from that of Refs. 3 and 8, the retarded time and radial coordinates, respectively, are being denoted here by u and r, with lower case ψ's for the Weyl tensor components. Indicating by primed letters those quantities defined in Refs. 3, 8, the translation to the present notation is achieved by

$$u' = 2^{\frac{1}{2}}u, \quad r' = 2^{-\frac{1}{2}}r, \quad l^{\mu'} = 2^{\frac{1}{2}}l^{\mu}, \quad m^{\mu'} = m^{\mu}, \quad n^{\mu'} = 2^{-\frac{1}{2}}n^{\mu};$$

thus, we have

$$\Psi_n' = 2^{\frac{1}{2}(2-n)}\psi_n, \quad \Psi_n^{m'} = 2^{-\frac{1}{2}(3-m)}\psi_n^m, \quad \dot{\Psi}_n^{m'} = 2^{-\frac{1}{2}(4+m)}\dot{\psi}_n^m,$$
$$\Phi_n' = 2^{\frac{1}{2}(1-n)}\phi_n, \quad \Phi_n^{m'} = 2^{-\frac{1}{2}(2+m)}\phi_n^m, \quad \dot{\Phi}_n^{m'} = 2^{-\frac{1}{2}(3+m)}\dot{\phi}_n^m,$$
$$\sigma' = 2^{\frac{1}{2}}\sigma, \quad \sigma^{0'} = 2^{-\frac{1}{2}}\sigma^0, \quad \dot{\sigma}^{0'} = \tfrac{1}{2}\dot{\sigma}^0, \quad \ddot{\sigma}^{0'} = 2^{-\frac{3}{2}}\ddot{\sigma}^0,$$
$$\rho' = 2^{\frac{1}{2}}\rho, \quad \tau' = \tau, \quad \alpha' = \alpha, \quad \alpha^{0'} = 2^{-\frac{1}{2}}\alpha^0, \quad \beta' = \beta,$$
$$\gamma' = 2^{-\frac{1}{2}}\gamma, \quad \lambda' = 2^{-\frac{1}{2}}\lambda, \quad \mu' = 2^{-\frac{1}{2}}\mu, \quad \nu' = \tfrac{1}{2}\nu,$$
$$\xi^{i'} = \xi^i, \quad X^{i'} = 2^{-\frac{1}{2}}X^i, \quad \omega' = 2^{-\frac{1}{2}}\omega, \quad U' = \tfrac{1}{2}U.$$

[5] E. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962); 4, 998 (1963).

[6] B. Kozarzewski, Acta Phys. Polon. 27, 775 (1965).

$$\psi_3 = \psi_3^0 r^{-2} + (\bar{\eth}\psi_2^0 + \bar{\phi}_1^0\phi_2^0)r^{-3} + O(r^{-4}), \tag{5}$$

$$\psi_4 = \psi_4^0 r^{-1} + (\bar{\eth}\psi_3^0)r^{-2} + O(r^{-3}), \tag{6}$$

$$\phi_1 = \phi_1^0 r^{-2} + \bar{\eth}\phi_0^0 r^{-3} + O(r^{-4}), \tag{7}$$

$$\phi_2 = \phi_2^0 r^{-1} + \bar{\eth}\phi_1^0 r^{-2} + O(r^{-3}), \tag{8}$$

$$\rho = -r^{-1} - \sigma^0\bar{\sigma}^0 r^{-3} + O(r^{-5}), \tag{9}$$

$$\sigma = \sigma^0 r^{-2} + (\bar{\sigma}^0\sigma^{0^2} - \tfrac{1}{2}\psi_0^0)r^{-4} + O(r^{-5}), \tag{10}$$

$$\alpha = \alpha^0 r^{-1} + \bar{\sigma}^0\bar{\alpha}^0 r^{-2} + \sigma^0\bar{\sigma}^0\alpha^0 r^{-3} + O(r^{-4}), \tag{11}$$

$$\beta = -\bar{\alpha}^0 r^{-1} - \sigma^0\alpha^0 r^{-2} - (\sigma^0\bar{\sigma}^0\bar{\alpha}^0 + \tfrac{1}{2}\psi_1^0)r^{-3} + O(r^{-4}), \tag{12}$$

$$\begin{aligned}\tau = {} & -\tfrac{1}{2}\psi_1^0 r^{-3} \\ & + \tfrac{1}{3}(\tfrac{1}{2}\sigma^0\bar{\psi}_1^0 - \bar{\eth}\psi_0^0 - 4\phi_0^0\bar{\phi}_1^0)r^{-4} + O(r^{-5}),\end{aligned} \tag{13}$$

$$\begin{aligned}\lambda = {} & \dot{\bar{\sigma}}^0 r^{-1} + \bar{\sigma}^0 r^{-2} \\ & + (\sigma^0\dot{\bar{\sigma}}^0\bar{\sigma}^0 + \tfrac{1}{2}\bar{\sigma}^0\psi_2^0 - \tfrac{1}{2}\bar{\phi}_0^0\phi_2^0)r^{-3} + O(r^{-4}),\end{aligned} \tag{14}$$

$$\begin{aligned}\mu = {} & -r^{-1} - (\sigma^0\dot{\bar{\sigma}}^0 + \psi_2^0)r^{-2} \\ & - (\sigma^0\bar{\sigma}^0 + \tfrac{1}{2}\bar{\eth}\psi_1^0 + \phi_1^0\bar{\phi}_1^0)r^{-3} + O(r^{-4}),\end{aligned} \tag{15}$$

$$\begin{aligned}\gamma = {} & -\tfrac{1}{2}\psi_2^0 r^{-2} \\ & - \tfrac{1}{6}(2\bar{\eth}\psi_1^0 - \alpha^0\psi_1^0 + \bar{\alpha}^0\bar{\psi}_1^0 + 6\phi_1^0\bar{\phi}_1^0)r^{-3} + O(r^{-4}),\end{aligned} \tag{16}$$

$$\nu = -\psi_3^0 r^{-1} - (\tfrac{1}{2}\bar{\eth}\psi_2^0 + \bar{\phi}_1^0\phi_2^0)r^{-2} + O(r^{-3}), \tag{17}$$

$$\begin{aligned}U = {} & -1 - \tfrac{1}{2}(\psi_2^0 + \bar{\psi}_2^0)r^{-1} \\ & - \tfrac{1}{6}(\bar{\eth}\psi_1^0 + \eth\bar{\psi}_1^0 + 6\phi_1^0\bar{\phi}_1^0)r^{-2} + O(r^{-3}),\end{aligned} \tag{18}$$

$$X^3 + iX^4 \sin\theta = \tfrac{1}{3}\psi_1^0 r^{-3} + O(r^{-4}), \tag{19}$$

X^3 and X^4 being real,

$$\xi^3 = r^{-1} - \sigma^0 r^{-2} + \sigma^0\bar{\sigma}^0 r^{-3} + O(r^{-4}), \tag{20a}$$

$$-i\xi^4 \sin\theta = r^{-1} + \sigma^0 r^{-2} + \sigma^0\bar{\sigma}^0 r^{-3} + O(r^{-4}), \tag{20b}$$

$$\omega = -\bar{\eth}\sigma^0 r^{-1} + (\sigma^0\eth\bar{\sigma}^0 - \tfrac{1}{2}\psi_1^0)r^{-2} + O(r^{-3}), \tag{21}$$

where

$$\alpha^0 = -\tfrac{1}{2}\cot\theta, \tag{22}$$

$$\psi_2^0 - \bar{\psi}_2^0 = \bar{\eth}^2\sigma^0 - \eth^2\bar{\sigma}^0 + \bar{\sigma}^0\dot{\sigma}^0 - \dot{\bar{\sigma}}^0\sigma^0, \tag{23}$$

$$\psi_3^0 = \eth\dot{\bar{\sigma}}^0, \tag{24}$$

$$\psi_4^0 = -\ddot{\bar{\sigma}}^0. \tag{25}$$

In the above, the "surfaces at infinity" ($r = \infty$, $u = \text{const}$) have been chosen to be spheres with the usual angular coordinates θ, φ. [Actually only Eqs. (19), (20), and (22) depend on the choice of angular coordinates.] The $\eth$ notation of Newman–Penrose[7] has also been used. Further terms could easily be calculated. The time development of the solution is determined (with the dot denoting $\partial/\partial u$) by

$$\dot{\psi}_0^0 + \eth\psi_1^0 - 3\sigma^0\psi_2^0 - 3\phi_0^0\bar{\phi}_2^0 = 0, \tag{26}$$

$$\begin{aligned}\dot{\psi}_0^1 - 4\bar{\eth}(\sigma^0\psi_1^0) + \bar{\eth}\eth\psi_0^0 + 4\bar{\phi}_1^0\eth\phi_0^0 & \\ -8\sigma^0\phi_1^0\bar{\phi}_1^0 - 4\phi_0^1\bar{\phi}_2^0 &= 0,\end{aligned} \tag{27}$$

$$\dot{\psi}_1^0 + \eth\psi_2^0 - 2\sigma^0\psi_3^0 - 2\phi_1^0\bar{\phi}_2^0 = 0, \tag{28}$$

$$\dot{\psi}_2^0 + \eth\psi_3^0 - \sigma^0\psi_4^0 - \phi_2^0\bar{\phi}_2^0 = 0, \tag{29}$$

$$\dot{\phi}_0^0 + \eth\phi_1^0 - \sigma^0\phi_2^0 = 0, \tag{30}$$

$$\dot{\phi}_0^1 + \bar{\eth}\eth\phi_0^0 - 2\bar{\eth}(\sigma^0\phi_1^0) = 0, \tag{31}$$

$$\dot{\phi}_1^0 + \eth\phi_2^0 = 0, \tag{32}$$

and the corresponding equations [see Eqs. (56), (57)] for the remaining $\dot{\psi}_0^m$, $\dot{\phi}_0^m$.

In these equations the quantities ψ_0^m, ϕ_0^m ($m = 0, 1, 2, \cdots$), ψ_1^0, $\psi_2^0 + \bar{\psi}_2^0$, ϕ_1^0, as functions of θ, φ, and the quantities σ^0 and ϕ_2^0 as functions of u, θ, φ are regarded as the given initial data with the remaining quantities expressed in terms of them. Thus, the data needed to determine a solution of the Einstein–Maxwell equations is the null-surface data for the Einstein vacuum equations together with the null-surface data for the flat-space Maxwell equations.[8] Also, the asymptotic symmetry group is just the Bondi–Metzner–Sachs (B.M.S.) group.[9,7]

We shall make use here of the results established in Ref. 2 on the spin weight of each of the ψ's and ϕ's, the expansion of a spin-weighted quantity in series of spin-s spherical harmonics[7] and some properties of $\eth$. In particular, the relations

$$\int {}_s\bar{Y}_{s,m}\bar{\eth}\eta\, dS = 0 \quad \text{and} \quad \int {}_sY_{s,m}\eth\phi\, dS = 0 \tag{33}$$

will be of great value here, where η is any quantity of spin weight $s + 1$ and ϕ is any quantity of spin weight $-s - 1$ defined on the (θ, φ) sphere and $dS = d\theta \sin\theta\, d\varphi$. [One way to prove Eq. (33) is to integrate by parts and use the fact[7] that $\eth_s\bar{Y}_{s,m} = 0$.]

As an example of the techniques to be used, the generalized Bondi–Sachs mass-loss theorem for the Einstein–Maxwell theory[10,11] can be easily obtained. The expression

$$M = -\int {}_0Y_{0,0}(\psi_2^0 + \sigma^0\dot{\bar{\sigma}}^0)\, dS \tag{34}$$

defines, up to a numerical factor, the Bondi–Sachs total mass[12,13] at retarded time u. (Of course, ${}_0Y_{0,0} = \frac{1}{2}\pi^{-\frac{1}{2}}$.) This M is real by Eq. (23) since

---

[7] E. Newman and R. Penrose, J. Math. Phys. 7, 863 (1966).

[8] A. Janis and E. Newman, J. Math. Phys. **6**, 902 (1965).

[9] R. K. Sachs, Phys. Rev. **128**, 2851 (1962).

[10] T. Morgan and J. Winicour, Bull. Am. Phys. Soc. **9**, 424 (1964).

[11] R. Penrose, Phys. Rev. Letters **10**, 66 (1963); in *Relativity, Groups and Topology*, C. DeWitt and B. DeWitt, Eds. (Gordon and Breach, Science Publishers, New York, 1964).

$$\int {}_0Y_{0,0}\eth^2\bar{\sigma}^0\, dS = 0.$$

From Eq. (29) we get

$$\dot{M} = -\int {}_0Y_{0,0}(\dot{\sigma}^0\dot{\bar{\sigma}}^0 + \phi_2^0\bar{\phi}_2^0)\, dS + \int {}_0Y_{0,0}\eth\psi_3^0\, dS. \quad (35)$$

The last term on the right vanishes (ψ_3^0 having spin weight -1), so if we integrate over u from u_0 to u_1 ($\mathrm{u}_1 > \mathrm{u}_0$) and use the fact that $\dot{\sigma}^0\dot{\bar{\sigma}}^0 + \phi_2^0\bar{\phi}_2^0$ is nonnegative, we obtain

$$M|_{\mathrm{u}_0}^{\mathrm{u}_1} \leq 0.$$

The Newman–Unti mass[3]

$$-\int {}_0Y_{0,0}\, \mathrm{Re}\, \psi_2^0\, dS$$

differs from M only when $\dot{\sigma}^0 \neq 0$, so in any transition between states with $\dot{\sigma}^0 = 0$ (e.g., between states stationary in the chosen coordinate system), this mass will likewise decrease.

III. CONSERVATION LAWS

Using Eqs. (31) and (32), we can easily obtain the conservation laws associated with the Maxwell field. To obtain the equation of charge conservation we multiply Eq. (32) by ${}_0Y_{0,0}$ and integrate over the sphere to obtain

$$\frac{d}{d\mathrm{u}}\int {}_0Y_{0,0}\phi_1^0\, dS = -\int {}_0Y_{0,0}\eth\phi_2^0\, dS. \quad (36)$$

By Eq. (33), the right side of this equation vanishes and we have

$$e \equiv \int {}_0Y_{0,0}\phi_1^0\, dS = \text{const.} \quad (37)$$

This is proportional to the charge ($+i$ × "magnetic charge"). A similar thing can now be done to Eq. (31), namely, multiply it by ${}_1\bar{Y}_{1,m}$ and integrate over the sphere to obtain

$$\frac{d}{d\mathrm{u}}\int {}_1\bar{Y}_{1,m}\phi_0^1\, dS = \int {}_1\bar{Y}_{1,m}\bar{\eth}A\, dS, \quad (38)$$

where $A \equiv -\eth\phi_0^0 + 2\sigma^0\phi_1^0$ is a quantity of spin weight 2. Again, by Eq. (33), the right side vanishes and we have

$$F_m \equiv \int {}_1\bar{Y}_{1,m}\phi_0^1\, dS = \text{const.} \quad (39)$$

The real and imaginary parts of F_{-1}, F_0, F_1 are the six conserved quantities for the Maxwell field described earlier,[1,2] which in flat space-time would be associated with the presence of incoming electromagnetic radiation.

One can obtain the ten conserved quantities associated with the gravitational field by considering Eq. (27), which we write as

$$\dot{\psi}_0^1 - 4\bar{\eth}(\sigma^0\psi_1^0) + \bar{\eth}\eth\psi_0^0 + 4R = 0, \quad (40)$$

where

$$R \equiv \bar{\phi}_1^0\eth\phi_0^0 - 2\sigma^0\phi_1^0\bar{\phi}_1^0 - \phi_0^1\bar{\phi}_2^0. \quad (41)$$

We first note that if ϕ is a quantity of nonnegative spin weight s, then a unique inverse to the operation $\bar{\eth}$ can be defined on this quantity if and only if[14]

$$\int {}_s\bar{Y}_{s,m}\phi\, dS = 0. \quad (42)$$

This can be seen by expanding ϕ in a series of spin-s spherical harmonics and using the properties of $\bar{\eth}$. Thus, one can uniquely define

$$\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) \quad \text{and} \quad \bar{\eth}^{-1}(\phi_0^1 - F),$$

where

$$\bar{E} \equiv \bar{e}_0Y_{0,0} \quad (43)$$

and

$$F \equiv \sum_{m=-1}^{+1} F_m\, {}_1Y_{1,m}. \quad (44)$$

The explicit expression for these quantities in an expansion in spin-s spherical harmonics is

$$\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) = -\sum_{l=1}^{\infty}\sum_{m=-l}^{l} \frac{\alpha_{lm}}{[l(l+1)]^{\frac{1}{2}}}\, {}_1Y_{l,m} \quad (45)$$

and

$$\bar{\eth}^{-1}(\phi_0^1 - F) = -\sum_{l=2}^{\infty}\sum_{m=-l}^{l} \frac{\beta_{lm}}{[(l-1)(l+2)]^{\frac{1}{2}}}\, {}_2Y_{l,m}, \quad (46)$$

where

$$\alpha_{lm} \equiv \int {}_0\bar{Y}_{l,m}\bar{\phi}_1^0\, dS$$

---

[12] H. Bondi, M. G. J. van der Burg, and A. W. K. Metzner, Proc. Roy. Soc. (London) **A269**, 21 (1962).

[13] R. K. Sachs, Proc. Roy. Soc. (London) **A270**, 103 (1962).

[14] More generally, for an *arbitrary* ϕ, we can define linear operators $\eth^\dagger$ and $\bar{\eth}^\dagger$—the "*generalized inverses*" of $\eth$, $\bar{\eth}$, respectively—by

$$\begin{aligned} \eth_s^\dagger Y_{l,m} &= (l-s+1)^{-\frac{1}{2}}(l+s)^{-\frac{1}{2}}_{s-1}Y_{l,m}, && \text{if } -s < l, \\ &= 0, && \text{if } -s = l, \\ \bar{\eth}_s^\dagger Y_{l,m} &= -(l-s)^{-\frac{1}{2}}(l+s+1)^{-\frac{1}{2}}_{s+1}Y_{l,m}, && \text{if } s < l, \\ &= 0, && \text{if } s = l. \end{aligned}$$

These satisfy $\eth\eth^\dagger\eth \equiv \eth$, $\eth^\dagger\eth\eth^\dagger \equiv \eth^\dagger$, $\eth^\dagger\eth \equiv \bar{\eth}\bar{\eth}^\dagger$, $\eth\eth^\dagger \equiv \bar{\eth}^\dagger\bar{\eth}$, $\bar{\eth}\bar{\eth}^\dagger\bar{\eth} \equiv \bar{\eth}$, $\bar{\eth}^\dagger\bar{\eth}\bar{\eth}^\dagger \equiv \bar{\eth}^\dagger$. [Compare R. Penrose, Proc. Cambridge Phil. Soc. **51**, 406 (1955), for the case of matrix generalized inverses.] The expressions $\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E})$ and $\bar{\eth}^{-1}(\phi_0^1 - F)$ used here can then be written concisely as $\bar{\eth}^\dagger\bar{\phi}_1^0$ and $\bar{\eth}^\dagger\phi_0^1$.

and

$$\beta_{lm} \equiv \int {}_1\bar{Y}_{l,m}\phi_0^1\, dS.$$

Using these results and also Eqs. (31), (32), and (41), we obtain

$$R = \frac{d}{du}[F\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) - \bar{\phi}_1^0\bar{\eth}^{-1}(\phi_0^1 - F)] - \bar{\eth}[\bar{\phi}_2^0\bar{\eth}^{-1}(\phi_0^1 - F)]. \quad (47)$$

Then Eq. (40) can be rewritten as

$$\frac{d}{du}[\psi_0^1 + 4F\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) - 4\bar{\phi}_1^0\bar{\eth}^{-1}(\phi_0^1 - F)] = \bar{\eth}Q, \quad (48)$$

where Q stands for an expression of spin weight 3. Hence, by Eq. (33), if we multiply by ${}_2\bar{Y}_{2,m}$ and integrate over the sphere, we obtain

$$G_m \equiv \int {}_2\bar{Y}_{2,m}[\psi_0^1 + 4F\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) - 4\bar{\phi}_1^0\bar{\eth}^{-1}(\phi_0^1 - F)]\, dS = \text{const.} \quad (49)$$

The real and imaginary parts of G_{-2}, G_{-1}, G_0, G_1, G_2 thus define ten conserved quantities for asymptotically flat Einstein–Maxwell fields. They are the same as the N–P gravitationally conserved quantities when the Maxwell field vanishes. Note that we can rewrite (49) in the alternative form

$$G_m \equiv \int {}_2\bar{Y}_{2,m}[\psi_0^1 + 4\phi_0^1\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) - 4\bar{E}\bar{\eth}^{-1}(\phi_0^1 - F)]\, dS; \quad (50)$$

the difference in the integrands in (49) and (50) being of the form ${}_2\bar{Y}_{2,m}\bar{\eth}(\bar{\eth}^{-1}A\bar{\eth}^{-1}B)$. If we add (49) to (50), we get a third, more symmetrical, expression

$$G_m \equiv \int {}_2\bar{Y}_{2,m}[\psi_0^1 + 2(\phi_0^1 + F)\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) - 2(\bar{\phi}_1^0 + \bar{E})\bar{\eth}^{-1}(\phi_0^1 - F)]\, dS. \quad (51)$$

The quantities G_m, F_m are invariant also under the *supertranslations* of the B.M.S. group. This does not follow from the argument given above, but the result can be obtained by use of an alternative technique.[15,2] The full details of this argument will be given elsewhere; in the case of F_m, the argument is given in its essentials in Ref. 2. For part of the argument of the next section, we shall, in fact, require this supertranslation invariance of the G_m and F_m, so we shall have to *assume* this result here.

We propose to establish a connection between the G_m, F_m and multipole moments of the fields in stationary states. This will then define "selection rules" for transitions between such states. Now, by an argument given in Ref. 2, we can show that, for any stationary system, allowable coordinates can be chosen so that $\sigma^0 = 0$ and for which there is no u dependence. Then we shall have $\phi_2^0 = 0$ [see Eq. (29)]. We can always obtain such coordinates by applying a B.M.S. transformation to the coordinates of any allowable N–U stationary system. (This may involve a supertranslation, however, so that the indentification of the F_m, G_m strictly requires their aforementioned supertranslation invariance.) Equations (31) and (30) now become

$$\bar{\eth}\eth\phi_0^0 = 0, \quad \text{whence} \quad \eth\phi_0^0 = 0,$$
$$\eth\phi_1^0 = 0. \quad (52)$$

These can be immediately integrated, yielding

$$\phi_0^0 = \sum_{m=-1}^{+1} A_m\, {}_1Y_{1,m},$$
$$\phi_1^0 = e\, {}_0Y_{0,0}. \quad (53)$$

Now, using this result, Eqs. (28), (27), and (26) become

$$\eth\psi_2^0 = 0,$$
$$\bar{\eth}\eth\psi_0^0 = 0, \quad \text{whence} \quad \eth\psi_0^0 = 0,$$
$$\eth\psi_1^0 = 0, \quad (54)$$

which integrate to

$$\psi_2^0 = -M\, {}_0Y_{0,0} = \bar{\psi}_2^0,$$
$$\psi_0^0 = \sum_{m=-2}^{+2} B_m\, {}_2Y_{2,m}, \quad (55)$$
$$\psi_1^0 = \sum_{m=-1}^{+1} C_m\, {}_1Y_{1,m}.$$

In order to determine the angular dependence of ψ_0^1 and ϕ_0^1, it is necessary to obtain the equations for the time development of ψ_0^2 and ϕ_0^2. This can be done by a straightforward but tedious calculation along the same lines as was used to obtain Eqs. (26) to (32), yielding

$$\dot{\psi}_0^2 = -\tfrac{1}{2}[\bar{\eth}\eth\psi_0^1 + 6\psi_0^1 + 15\psi_0^0\psi_2^0 - 10\psi_1^0\psi_1^0 + 6\bar{\phi}_1^0\eth\phi_0^1 - 2\bar{\alpha}^0\bar{\phi}_0^0\eth\phi_0^0] + \Omega, \quad (56)$$

where Ω is a quantity that vanishes when

$$\sigma^0 = \phi_2^0 = 0$$

and

$$\dot{\phi}_0^2 \doteq -\tfrac{1}{2}[\bar{\eth}\eth\phi_0^1 + 4\phi_0^1 + 3(\psi_2^0 + \bar{\psi}_2^0)\phi_0^0 - 4\psi_1^0\phi_1^0] + \Pi, \quad (57)$$

where Π is a quantity that vanishes when $\sigma^0 = \phi_0^2 = 0$.

---

[15] R. Penrose, in *Relativity Theory and Astrophysics*, J. Ehlers, Ed. (American Mathematical Society, Providence, R.I., 1967), Vol. VIII.

In the stationary case with the above choice of coordinates, Eq. (57) yields

$$\bar{\eth}\eth\phi_0^1 + 4\phi_0^1 = 4\psi_1^0\phi_1^0 - 6\psi_2^0\phi_0^0. \tag{58}$$

This can be integrated by noting that the right side can be expressed as a sum over ${}_1Y_{1,m}$ and then using the properties of $\bar{\eth}\eth$ on a quantity of spin weight 1 to obtain

$$\phi_0^1 = \sum_{m=-2}^{+2} D_m\, {}_1Y_{2,m} + H, \tag{59}$$

where

$$H = \phi_1^0\psi_1^0 - \tfrac{3}{2}\psi_2^0\phi_0^0.$$

The Maxwell conserved quantities are

$$F_m = \int {}_1Y_{1,m} H\, dS \tag{60}$$

[and, in fact, $H = F$; see (44)].

The e, C_m, M, and A_m are, respectively, proportional to the electric ($+i$ × magnetic) charge, mass dipole moment ($+i$ × angular momentum), mass, and electric ($+i$ × magnetic) dipole moment. Thus for stationary fields, the Maxwell conserved quantities are explicit quadratic expressions in these 4 moments. (In fact, essentially: mass × electric dipole − charge × gravitational dipole; this being an origin-independent expression.)

Finally, to obtain an expression for ψ_0^1 in stationary states, we use Eq. (56) (noting that $\eth\phi_0^0 = 0$) to obtain

$$\bar{\eth}\eth\psi_0^1 + 6\psi_0^1 = 10(\psi_1^0)^2 - 15\psi_0^0\psi_2^0 - 6\bar{\phi}_1^0\eth\phi_0^1. \tag{61}$$

This can be integrated by noting that the right side can be expressed as a sum over ${}_2Y_{2,m}$ and using the properties of $\bar{\eth}\eth$ to obtain

$$\psi_0^1 = \sum_m J_m\, {}_2Y_{3,m} + \tfrac{5}{3}(\psi_1^0)^2 - \tfrac{5}{2}\psi_0^0\psi_2^0 - \bar{\phi}_1^0\eth\phi_0^1. \tag{62}$$

We can now use these expressions to evaluate the conserved quantities in stationary states. From Eqs. (45) and (46) we have that in stationary states

$$\bar{\eth}^{-1}(\bar{\phi}_1^0 - \bar{E}) = 0$$

and

$$\bar{\eth}^{-1}(\phi_0^1 - F) = -\tfrac{1}{2}\sum_m D_m\, {}_2Y_{2,m} = -\tfrac{1}{4}\eth\phi_0^1 \tag{63}$$

and, thus, Eq. (51) becomes

$$G_m = \int {}_2\bar{Y}_{2,m}\{\psi_0^1 + \bar{\phi}_1^0\eth\phi_0^1\}\, dS, \tag{64}$$

which yields by Eq. (62)

$$G_m = \int {}_2\bar{Y}_{2,m} K\, dS, \tag{65}$$

where

$$K = \tfrac{5}{3}(\psi_1^0)^2 - \tfrac{5}{2}\psi_0^0\psi_2^0.$$

Thus, in stationary states, the conserved quantities can be expressed as exactly the same combination of the mass, dipole moment, and quadrupole moment as in the gravitational case. This means that the implied selection rules for the transitions between stationary states of the Einstein–Maxwell field are the same as those between stationary states of the vacuum gravitational field but with additional selection rules imposed by the Maxwell conserved quantities [Eq. (60)]. Whether or not transitions between exactly stationary states[16] are possible at all, however, even when K and H are both unaltered, remains an open question at present.

---

[16] It seems unlikely that transitions between two different states can ever be achieved if each is to be without time dependence and with $\sigma^0 = 0$ in the *same* coordinate system. This is indicated by some work of Unti on axially symmetric vacuum fields. Thus, it is important that we have supertranslation invariance for the F_m and G_m, for the above selection rules to have any content.

42

THE SINGULARITIES OF GRAVITATIONAL COLLAPSE AND COSMOLOGY

Proc. Roy. Soc. London Ser. A **314(1519)** (1970), 529–48. With S. W. Hawking

Following my demonstration of the singularity theorem put forward in Chapter 20, Stephen Hawking produced a series of papers (published mainly in the Royal Society's Proceedings) containing technical mathematical advances which effectively generalized this result so that the implication of the presence of space-time singularities in generic situations would apply also to the cosmological context, where certain restrictive assumptions that I had made—primarily the existence of a global non-compact Cauchy hypesurface—could be removed. Stimulated by my preparations for the Battelle lectures of Chapter 36, which included a rapid presentation of Stephen's main results, I realized that some improvements could be made leading to a more powerful theorem, and Stephen independently came to similar conclusions. We combined forces to produce the results described here.

The singularities of gravitational collapse and cosmology

BY S. W. HAWKING
Institute of Theoretical Astronomy, University of Cambridge

AND R. PENROSE
Department of Mathematics, Birkbeck College, London

(*Communicated by H. Bondi, F.R.S.—Received* 30 *April* 1969)

A new theorem on space-time singularities is presented which largely incorporates and generalizes the previously known results. The theorem implies that space-time singularities are to be expected if *either* the universe is spatially closed *or* there is an 'object' undergoing relativistic gravitational collapse (existence of a trapped surface) *or* there is a point p whose past null cone encounters sufficient matter that the divergence of the null rays through p changes sign somewhere to the past of p (i.e. there is a minimum apparent solid angle, as viewed from p for small objects of given size). The theorem applies if the following four physical assumptions are made: (i) Einstein's equations hold (with zero or negative cosmological constant), (ii) the energy density is nowhere less than minus each principal pressure nor less than minus the sum of the three principal pressures (the 'energy condition'), (iii) there are no closed timelike curves, (iv) every timelike or null geodesic enters a region where the curvature is not specially alined with the geodesic. (This last condition would hold in any sufficiently general physically realistic model.) In common with earlier results, timelike or null geodesic incompleteness is used here as the indication of the presence of space-time singularities. No assumption concerning existence of a global Cauchy hypersurface is required for the present theorem.

1. INTRODUCTION

An important feature of gravitation, for very large concentrations of mass, is that it is essentially *unstable*. This is due, in the first instance, to its r^{-2} attractive character. But, in addition, when general relativity begins to play a significant role, other instabilities may also arise (cf. Chandrasekhar 1964). The instability of gravitation is not manifest under normal conditions owing to the extreme smallness of the gravitational constant. The pull of gravity is readily counteracted by other forces. However, this instability *does* play an important dynamical role when large enough concentrations of mass are present. In particular, as the work of Chandrasekhar (1935) showed, a star of mass greater than about 1.3 times that of the Sun, which has exhausted its resources of thermal and nuclear energy, cannot sustain itself against its own gravitational pull, so a *gravitational collapse* ensues. It has sometimes been suggested also that, on a somewhat larger scale, some form of gravitational collapse may be taking place in quasars, or perhaps in the centres of (some?) galaxies. Finally, on the scale of the universe as a whole, this instability shows up again in those models for which the expansion eventually reverses, and the entire universe becomes involved in a gravitational collapse. In the reverse direction in time there is also the 'big bang' initial phase which is common to most relativistic expanding models. This again may be regarded as a manifestation of the instability of gravitation (in reverse).

But what is the ultimate fate of a system in gravitational collapse? Is the picture that is presented by symmetrical exact models accurate, according to which a *singularity in space-time* would ensue? Or may it not be that any asymmetries present might cause the different parts of the collapsing material to miss each other, so possibly to lead to some form of *bounce*? It seems that until comparatively recently many people had believed that such an asymmetrical bounce might indeed be possible to achieve, in a manner consistent with general relativity (cf. particularly, Lindquist & Wheeler 1957; Lifshitz & Khalatnikov 1963). However, some recent theorems† (Penrose 1965*a*; Hawking 1966*a*, *b*; **H**; Geroch 1966) have ruled out a large number of possibilities of this kind. The present paper carries these results further, and considerably strengthens the implication that a singularity-free bounce (of the type required) does not seem to be realizable within the framework of general relativity.

In the first theorem (referred to as I; see Penrose 1965*a*; cf. also Penrose 1966; **P**; Hawking 1966*c*) the concept of the existence of a *trapped surface*‡ was used as a characterization of a gravitational collapse which has passed a 'point of no return'. On the basis of a *weak energy condition*,‡ the intention was to establish the existence of space-time singularities from the existence of a trapped surface. Unfortunately, however, theorem I required, as an additional hypothesis, the existence of a non-compact global Cauchy hypersurface. Although 'reasonable' from the point of view of classical Laplacian determinism, the assumption of the existence of a global Cauchy hypersurface is hard to justify from the standpoint of general relativity. Also, it is violated in a number of exact models. Furthermore, the non-compactness assumption used in theorem I applies only if the universe is 'open'.

The second theorem (Hawking 1966*a*), and its improved version (referred to as II, see **H**; cf. also Hawking (1966*c*) and **P**), required the existence of a compact spacelike hypersurface with everywhere diverging normals. Thus it applies to 'closed', everywhere expanding, universe models. For such models II implies the existence of an initial (e.g. 'big bang' type) singularity. However, this condition on the normals may well not be applicable to the actual universe (particularly if there are local collapsing regions), even if the universe *is* 'closed'. Also, the condition is virtually unverifiable by observation.

The third and fourth results (referred to as III and IV; see Geroch (1966) and Hawking (1966*b*), respectively) again apply to 'closed' universe models (i,e. containing a compact, spacelike hypersurface), but which do not have to be assumed to be everywhere expanding. However, III required the somewhat unnatural assumption of the non-existence of 'horizons', while IV required that the given compact hypersurface be a global Cauchy hypersurface. Thus, III and IV could be objected to on grounds similar to those of I.

---

† We use **H** for referring to Hawking (1967) and **P** for referring to Penrose (1968).

‡ The precise meanings of these terms will be given in §3.

The fifth theorem (referred to as V; see **H**, also Hawking (1966*c*) and **P**) does not suffer from objections of this kind, but the requirement on which it was based—namely that the divergence of all timelike and null geodesics through some point p changes sign somewhere to the past of p—is somewhat stronger than one would wish. Theorem V would be considerably more useful in application if the above requirement referred only to *null* geodesics.

In this paper we establish a new theorem, which, with two reservations, effectively incorporates all of I, II, III, IV and V while avoiding each of the above objections. In its physical implications, our theorem falls short of completely superseding these previous results only in the following two main respects. In the first instance we shall require the non-existence of closed time like curves. Theorem II (and II alone) did not require such an assumption. Secondly, in common with II, III, IV and V, we shall require the slightly stronger energy condition given in (3.4), than that used in I. This means that our theorem cannot be directly applied when a *positive cosmological constant* λ is present. However, in a collapse, or 'big bang', situation we expect large curvatures to occur, and the larger the curvatures present the smaller is the significance of the value of λ. Thus, it is hard to imagine that the value of λ should qualitively affect the singularity discussion, except in regions where curvatures are still small enough to be comparable with λ. We may take I as a further indication (though not a proof) of this. In a similar way, II may be taken as a strong indication that the development of closed timelike curves is not the 'answer' to the singularity problem. Of course, such causality violation would carry with it other very serious problems, in any case.

The energy condition (3.4) used here (and in II, III, IV and V) has a very direct physical interpretation. It states, in effect, that 'gravitation is always attractive' (in the sense that neighbouring geodesics near any one point accelerate, on the average, towards each other). Our theorem will apply, in fact, in theories other than classical general relativity provided gravitation remains attractive. In particular, we can apply our results in the theory of Brans & Dicke (1961), using the metric for which the field equations resemble Einstein's (cf. Dicke 1962). The gravitational constant could, in principle, change sign in this theory, but only via a region at which it becomes infinite. Such a region could reasonably be called a 'singularity' in any case. On the other hand, gravitation does not always remain attractive in the theory of Hoyle & Narliker (1963) (owing to the effective negative energy of the C-field) so our theorem is not directly applicable in this theory. We note, finally, that in Einstein's theory (with 'reasonable' sources) it is only $\lambda > 0$ which can prevent gravitation from being always attractive, the λ term representing a 'cosmic repulsion'.

In common with all the previous results I, . . . , V, our theorem will not give very much information as to the nature of the space-time singularities that are to be inferred on the basis of Einstein's theory. If we accept that 'causality breakdown' is unlikely to occur (because of philosophical difficulties encountered with closed timelike curves and because theorem II suggests that such curves probably do not help in the singularity problem in any case), then we are led to the view that the instability of gravitation presumably† results in regions of enormously large curvature occurring in our universe. These curvatures would have to be so large that our present concepts of local physics would become drastically modified. While the quantum effects of gravitation are normally thought to be significant only when curvatures approach 10^{33} cm^{-1}, all our local physics is based on the Poincaré group being a good approximation of a local symmetry group at dimensions greater than 10^{-13} cm. Thus, if curvatures ever even approach 10^{13} cm^{-1}, there can be little doubt but that extraordinary local effects are likely to take place.

When a singularity results from a collapse situation in which a trapped surface has developed, then any such local effects would not be observable outside the collapse region. It is an open question whether physically realistic collapse situations, resulting in singularities, will sometimes arise *without* trapped surfaces developing (cf. Penrose 1969). If they do, it is likely that such singularities could (in principle) be observed from outside. Of course, the initial 'big bang' singularity of the Robertson–Walker models is an example of a singularity of the observable type. However, our theorem yields no information as to the observability of singularities in general. We cannot even rigorously infer whether the implied singularities are to be expected in the 'past' or the 'future'. (In this respect our present theorem yields somewhat less information than I, II, or V.)

Our theorem will be directly applicable to *any one* of the following three situations. First, to the existence of a trapped surface; secondly, to the existence of of a compact space-like hypersurface; thirdly, to the existence of a point whose null-cone begins to 'converge again' somewhere to the past of the point. We assume the energy condition and the non-existence of closed timelike curves. On the basis of this (and another very minor assumption which merely rules out some highly special models) we deduce that singularities will develop in fully general situations involving a collapsing star, *or* in a spatially closed universe, *or* (taking the point in question in the third case to be the earth at the present time) if the apparent solid angle subtended by an object of a given intrinsic size reaches some minimum when the object is at a certain distance from us. We show, in an appendix, that this last condition is indeed likely to be satisfied in our universe, assuming the correctness of the normal interpretation of the 2.7 K background radiation. A similar discussion was given earlier by Hawking & Ellis (1968) in connexion with theorem V. Since we now have a stronger theorem, we can use somewhat weaker physical assumptions concerning the radiation.

In §2 we give a number of lemmas and definitions that will be needed for our theorem. The precise statement of the theorem will be given in § 3. This statement

---

† We must always bear in mind that a local 'energy-condition' (cf. (3.4)) is being assumed here, which might be violated not only in a modified Einstein theory (e.g. 'C-field'), but also in the standard theory if we were allowed to have very 'peculiar' matter under extreme conditions. The quantum field-theoretic requirement of positive-definiteness of energy (in order that the vacuum remain stable) is of great relevance here, but its status is perhaps not completely clear (cf. Sexl & Urbantke 1967 for example).

is presented in a rather general form, which is somewhat removed from the actual applications. The main applications are given in a corollary to the theorem. One slight advantage of the form of statement that we have chosen will be that it enables a small amount of information to be extracted about the actual nature of the singularities. This is that (at least) one timelike or null geodesic must enter (or leave) the singularity not only in a finite proper (or affine) time, but also in such a way that none of the neighbouring initially parallel geodesics has time to be focused towards it before the singularity is encountered.

2. DEFINITIONS AND LEMMAS

A four-dimensional differentiable (Hausdorff and paracompact†) manifold M will be called a *space-time* if it possesses a pseudo-Riemannian metric of hyperbolic normal signature $(+, -, -, -)$ and a time-orientation. (In fact the following arguments will apply equally well if M has any dimension $\geqslant 3$; also, the time-orientability of M need not really be assumed if we are prepared to apply the arguments to a twofold covering of M.) There will be no real loss of generality in physical applications if we assume that M and its metric are both C^∞. However, the arguments we use actually only require the metric to be C^2.

We shall be concerned with *timelike curves* and *causal curves* on M. (When we speak of a 'curve', we shall, according to context, mean either a continuous map into M of a connected closed portion of the real line, or else the image in M of such a map.) For definiteness we choose our timelike curves to be *smooth*, with future-directed tangent vectors everywhere strictly timelike, including at its end-points. A causal curve is a curve obtainable as a limiting case of timelike curves‡ (cf. Siefert 1967; Carter 1967); it is continuous but not necessarily everywhere smooth; where smooth, its tangent vectors are either timelike or null. A timelike or causal curve will require end-points if it can be extended as a causal curve either into the past or the future (cf. **P**, p. 187). If it continues indefinitely into the past [*resp.* future] it will be called *past-inextendible* [*resp. future-inextendible*]. If both past- and future-inextendible it is called *inextendible*.

If $p, q \in M$, we write $p \ll q$ if there is a timelike curve with past end-point p and future end-point q; we write $p \prec q$ if either $p = q$ or there is a causal curve from p to q (cf. Kronheimer & Penrose 1967). If $p \prec q$ but not $p \ll q$, then there is a null geodesic from p to q, or else $p = q$. If $p \ll q$ and $q \prec r$, or if $p \prec q$ and $q \ll r$, then $p \ll r$. We do not have $p \ll p$ unless M contains *closed timelike curves*. A subset of M is called *achronal* if it contains no pair of points p, q with $p \ll q$.

---

† Geroch (1968*b*) has shown that the assumption of paracompactness is not actually necessary for a space-time, being a consequence of the other assumptions for a space-time manifold.

‡ Except for very minor parts of our discussion, the fact that we are allowing our causal curves not to be smooth plays no significant role in this paper, but it is useful for the general theory. A continuous map of the connected closed interval $\Gamma \subset \mathfrak{R}$, into M, can be characterized as a causal curve by the fact that if $[a, b] \in \Gamma$ and if A, B and C are neighbourhoods in M of the images of a, b and $[a, b]$, respectively, then there exists a timelike curve lying in C with one end-point in A and another end-point in B.

We shall, for the most part, use terminology, definitions and some basic results as given in **P**. (However we use 'causal' for curves referred to in **P** as 'nonspacelike' and 'achronal' for sets referred to in **P** as 'semispacelike'; cf. Carter 1967.) As in Kronheimer & Penrose (1967), we write $I^+(p)$ for the open future of a point $p \in M$, i.e. $I^+(p) = \{x : p \ll x\}$ and $I^+[S]$ for the open future of a set $S \subset M$, i.e. $I^+[S] = \bigcup_{p \in S} I^+(p)$. (The sets $I^+[S]$ are open in the manifold topology for M.) Similarly, $J^+(p) = \{x : p \prec x\}$; $J^+[S] = \bigcup_{p \in S} J^+(p)$. These are not always closed sets.) We define

$$E^+(S) = J^+[S] - I^+[S]. \tag{2.1}$$

Then $E^+(S)$ is *part* of the *boundary* $\dot{I}^+[S]$ of $I^+[S]$ but not necessarily all of it. The sets $I^-(p)$, $I^-[S]$, $J^-(p)$, $J^-[S]$ and $E^-(S)$ are defined similarly, but with future and past interchanged.

For any set $S \in M$ we can define the (*future*) *domain of dependence* $D^+(S)$ and *Cauchy horizon* $H^+(S)$ by

$$D^+(S) = \{x : \text{every past-inextendible timelike curve through } x \text{ meets } S\} \tag{2.2}$$

and

$$\begin{aligned} H^+(S) &= \{x : x \in D^+(S),\ I^+(x) \cap D^+(S) = \varnothing\} \\ &= D^+(S) - I^-[D^+(S)]. \end{aligned} \tag{2.3}$$

The sets $D^-(S)$ and $H^-(S)$ are correspondingly defined. (These definitions are chosen to agree with **P**; they differ somewhat from those of **H**.) We shall be concerned only with the cases when S is an *achronal closed set*. Then $D^+(S)$ is a closed set and $H^+(S)$ is an *achronal closed set*. One easily verifies:

$$I^+[H^+(S)] = I^+[S] - D^+(S). \tag{2.4}$$

Define the *edge* of an achronal closed set S to be the set of points $p \in S$ such that† if $r \ll p \ll q$, with γ a timelike curve from r to q, containing p, then every neighbourhood of γ contains a timelike curve from r to q not meeting S. It follows that edge (S) is in fact the set of points in whose vicinity S fails to be a C°—manifold (S achronal and closed). We have (cf. **P**, p. 191) edge $(S) \subset H^+(S)$. (In fact edge(S) = edge $(H^+(S))$.) Furthermore:

LEMMA (2.5). *Every point of* $H^+(S) -$ edge(S) *is the future end-point of a null geodesic on* $H^+(S)$ *which can be extended into the past on* $H^+(S)$ *either indefinitely, or until it meets* edge(S).

For the proof, see **P**, p. 217 (compare **H**).

A similar result (which follows at once from **P**, p. 216; **H**) is (with S closed and achronal).

LEMMA (2.6). *Every point* $p \in \dot{I}^+[S] - S$ *is the future end-point of a null geodesic on* $\dot{I}^+[S]$ *which can be extended into the past on* $\dot{I}^+[S]$ *either indefinitely (if* $p \in \dot{I}^+[S] - E^+(S)$*) or until it meets edge*(S) *(whence* $p \in E^+(S)$*).*

We say that *strong causality* holds at p if arbitrarily small neighbourhoods of p exist, each intersecting no timelike curve in a disconnected set. (Roughly speaking, this means that timelike curves cannot leave the vicinity of p and then return to it; i.e. M does not 'almost' contain closed timelike curves.) We must say 'arbitrarily small', rather than 'every', in the above definition because of the existence of 'hour-glass shaped' (or even 'ball shaped') neighbourhoods of any point in *any* space-time, which are left and re-entered by a timelike curve. To avoid this feature, let us call an open set Q *causally convex* (**P**, p. 224) if Q intersects no timelike curve in a disconnected set. Thus, strong causality holds at p if and only if p possesses arbitrarily small causally convex neighbourhoods (in which case, the 'Alexandrov neighbourhoods' $I^+(q) \cap I^-(r)$ will suffice, with $q \ll p \ll r$). A causally convex open set which lies inside a convex normal coordinate ball with compact closure† will be called a *local causality neighbourhood* (**H**, p. 192). Strong causality holds at every point of a local causality neighbourhood. The only properties of a local causality neighbourhood that we shall in fact use, are that it is open and causally convex, that it contains no past- (or future) -inextendible null geodesic and that any point at which strong causality holds possesses such a neighbourhood.

---

† This replaces the definition of edge (S) given in **P**, which was not quite correctly stated.

A property of $D^+(S)$ we shall require is the following. Again, S is to be achronal and closed.

LEMMA (2.7). *If* $p \in \operatorname{int} D^+(S)$, *then* $J^-(p) \cap J^+[S]$ *is compact.*

This follows from **H**. (See also **P**, p. 227: if edge$(S) = \varnothing$, and strong causality holds at each point‡ of S, we have the stronger result that int $D^+(S)$ is precisely the set of $p \in I^+[S]$ for which $J^-(p) \cap J^+[S]$ is both compact and contains no point at which strong causality fails. Lemma (2.7) follows by similar reasoning.)

We shall require the concept of *conjugate points* on a causal (i.e. timelike or null) geodesic. Two points p and q on a causal geodesic γ are said to be *conjugate* if a geodesic 'neighbouring' to γ 'meets' γ at p and at q. Somewhat more precisely, the congruence of geodesics through p in the neighbourhood of γ has q as a *focal* point, that is, a point where the divergence of the congruence becomes infinite. (This focal point will in general be an 'astigmatic' focal point. It is a point of the 'caustic' of the congruence. Precise definitions of conjugate points will be found in Milnor (1963), Hicks (1965), Hawking (1966*a*).) The relation of conjugacy is symmetrical in p and q. The above definition still holds if the roles of p and q are reversed. The property of conjugate points that we shall require is the following (for the timelike case, see Boyer (1964), Hawking (1966*a*, *c*), cf. Milnor (1963); for the null case see Hawking (1966*c*) and also **P**, p. 215, for an equivalent result).

LEMMA (2.8). *If a causal geodesic* γ *from* p *to* q *contains a pair of conjugate points between* p *and* q, *then there exists a timelike curve from* p *to* q *whose length exceeds that of* γ.

We use the term 'length' for a causal curve to denote its proper time integral. A timelike geodesic is *locally* a curve of maximum length. As a corollary of lemma (2.8) we have:

---

† This condition was not explicitly included in the definition given in **H**.

‡ This condition should have been included in the conditions on $\mathscr{H}$ in lemma V of **P**.

LEMMA (2.9). *If γ is a null geodesic lying on $I^+[S]$ or on $H^+(S)$ for some $S \subset M$, then γ cannot contain a pair of conjugate points except possibly at its end-points.*

Another consequence of lemma (2.8) is the following result:

LEMMA (2.10). *If M contains no closed timelike curves and if every inextendible null geodesic in M possesses a pair of conjugate points, then strong causality holds throughout M.*

Proof. The result has been given in Hawking (1966*c*). We repeat the argument here since this reference is not readily available. Suppose strong causality fails at p. Let B be a normal coordinate neighbourhood of p and Q_i a nested sequence of neighbourhoods of p converging on p. Now there is a timelike curve originating in Q_i which leaves B at a point $q_i \in \dot{B}$, re-enters B and returns to Q_i. As $i \to \infty$ the q_i have an accumulation point q on $\dot{B}$ ($\dot{B}$ being compact). The geodesic pq in B cannot be timelike (since otherwise $I^-(q)$ would contain some Q_i, so closed timelike curves would result), nor spacelike. It must therefore be null. Furthermore, strong causality must also fail at q. Repeating the argument with q in place of p, we obtain a new null geodesic qr. In fact this must be the continuation of pq, since otherwise closed timelike curves would result. Continuing the process indefinitely both into the future and into the past we get an inextendible null geodesic γ at every point of which strong causality must fail. By hypothesis γ contains a pair of conjugate points. Thus by lemma (2.8) two of its points can be connected by a timelike curve. It follows that each point of some neighbourhood of one of these point can be joined by a timelike curve to each point of some neighbourhood of the other. This leads at once to the existence of closed timelike curves (because of strong causality violation), contrary to hypothesis. This establishes the lemma.

An important consequence of strong causality is the following result.

LEMMA (2.11). *Let $p \ll q$ be such that the set $J^+(p) \cap J^-(q)$ is compact and contains no points at which strong causality fails. Then there is a timelike geodesic from p to q which attains the maximum length for timelike curves connecting p to q.*

This result was proved by Siefert (1967). The result is, in effect, also contained in the earlier work of Avez (1963). (Unfortunately Avez's analysis contains some errors owing to the fact that the possibility of strong causality breakdown is not duly taken into account.) Lemma (2.11) follows also from lemma V in **P** (p. 227) in conjunction with VI of **P** (p. 228), as applied to the closed achronal set $\dot{I}^-(q)$. In fact, lemma (2.11) can be generalized: if C is a compact subset of M containing no points at which strong causality fails, then the maximum length for all timelike curves contained in C is *attained* (though not necessarily by a geodesic). The essential feature of this situation is that the *space of causal curves* contained in C is compact, the length of a causal curve being an upper semi-continuous function of the curve. For this, we need the appropriate topology on the space of causal curves. (See Seifert (1967); cf. also Avez (1963)). But it will not be necessary to enter into the general discussion here, as lemma (21.1) is all we shall need.

We define a *future-trapped* [*resp. past-trapped*] set to be a non-empty achronal closed† set $S \subset M$ for which $E^+(S)$ [*resp.* $E^-(S)$] is compact. (Note that $E^+(S)$ [*resp.* $E^-(S)$] must then be a closed achronal set.) Any future-trapped set S must itself be compact, since $S \subset E^+(S)$.) An example of a future-trapped set is illustrated in figure 1. We now come to our main lemma.

Lemma (2.12). *If S is a future-trapped set for which strong causality holds at every point of $\bar{I}^+[S]$, then there exists a future-inextendible timelike curve $\gamma \subset \operatorname{int} D^+(E^+(S))$.*

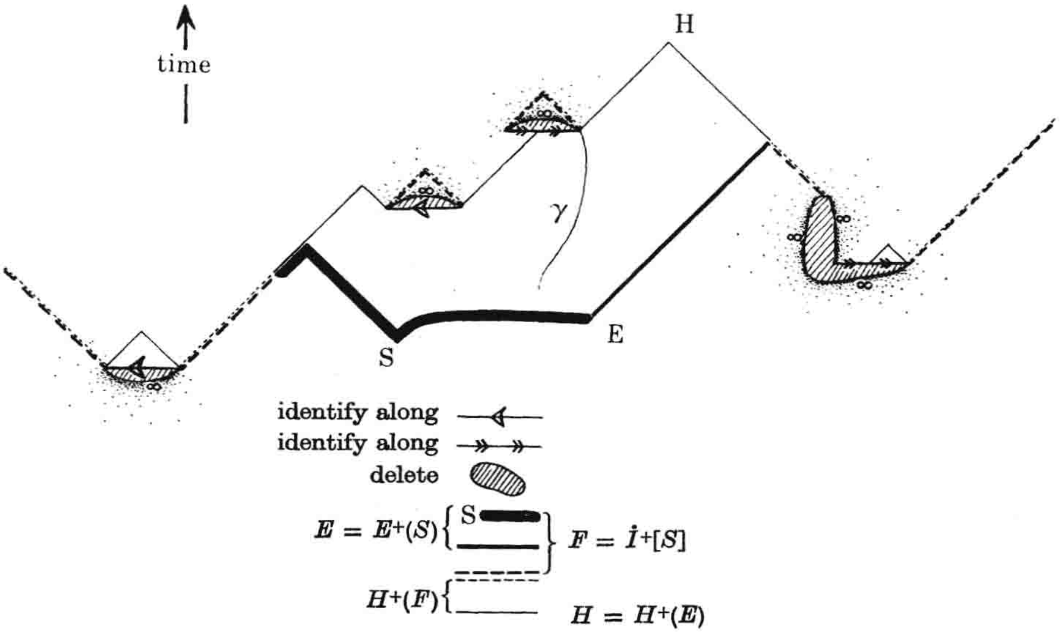


Figure 1. A future-trapped set S, together with the associated achronal sets $E = E^+(S)$, $F = \dot{I}^+[S]$, $H^+(F)$, $H = H^+(E)$. (For the proof of lemma (2.12).) The figure is drawn according to the conventions whereby null lines are inclined at 45°. The diagonally shaded portions are excluded from the space-time and some identifications are made. The symbol ∞ indicates regions 'at infinity' with respect to the metric. A future-inextendible timelike curve $\gamma \in D^+(E)$ is depicted, in agreement with the conclusion of lemma (2.12).

Proof.‡ We first make some remarks concerning the relation between $E = E^+(S)$ and $F = \dot{I}^+[S] = \dot{I}^+[E]$, and between their domains of dependence and their Cauchy horizons. We have $E \subset F$, whence $D^+(E) \subset D^+(F)$. We have $\operatorname{edge}(F) = \varnothing$, so it follows from lemma (2.5) that each point of $F - E$ lies on a past-inextendible null geodesic on $F - E$. (These null geodesics extend into the future, while remaining on F–E, perhaps reaching a future end-point on edge(E). We readily obtain $D^+(F) - D^+(E) = H^+(F) - H^+(E) = F - E$, so int $D^+(E) =$ int $D^+(F)$.)

---

† The condition that S be closed could be omitted from this definition if desired. For, if S is achronal with $E^+(S)$ compact, then $E^+(S) = E^+(\bar{S})$. Another apparent weakening of the definition of 'future-trapped' for a closed achronal non-empty set S would be to say that $E^+(S)$ has compact closure. ($E^+(S)$ is not always a closed set, for general S.) This definition would be equivalent to the one we use, provided strong causality holds.

‡ This argument follows, to some extent, one given in **H** (pp. 198–9). It may also serve as a replacement for the final argument given in **P** (on p. 230) which was not stated correctly.

We shall show that $H = H^+(E)$ is non-compact or empty. For, suppose H is compact. Then we can cover H with a finite number of local causality neighbourhoods B_i. If H is non-empty, then $D^+(E) \not\supset I^+[S]$. Let $p \in I^+[S] - D^+(E)$ with p near H and suppose $p \in B_k$. Since $p \in I^+[S]$, a timelike curve η exists connecting S to p. Since $p \notin D^+(E)$, it follows that η meets H at a point p_0, say. We wish to construct a point $q \in I^+[S] - D^+(E)$ with $q \ll p$, $q \notin B_k$ and $q \in B_l$, say. If $p_0 \notin \bar{B}_k$ we can achieve this by taking q just to the future of p_0 on η. If $p_0 \in \bar{B}_k$ we follow the past-inextendible null geodesic ζ through p_0 on $H^+(F)$ (cf. (2.5)). Now ζ must leave $\bar{B}_k$ (since $\bar{B}_k$ is compact) and so contains a point $p_1 \notin \bar{B}_k$ on $H^+(F)$. We have $p_1 \prec p_0 \ll p$, so $p_1 \ll p$. Choosing q near p_1, with $p_1 \ll q \ll p$, we have $q \notin B_k$ and $q \in B_l$, say, where $q \in I^+(p_1) \subset I^+[H^+(F)] = I^+[S] - D^+(E)$ as required (cf. (2.4)). Repeating the procedure, we can find $r \in I^+[S] - D^+(E)$ with $r \ll q$, $r \notin B_l$ and $r \in B_m$, say, etc. Since the B_i are finite in number, there must be two of $p, q, r, \ldots$, in the same B_i, hence violating causal convexity. Thus, H if non-empty, must be non-compact, as required.

Now by a well known theorem (cf. Steenrod 1951, p. 201) we can choose a smooth (future-directed) timelike vector field on M. Form the integral curves $\{\mu\}$ of this vector field. Then each μ which meets H must also meet E (since $H \subset D^+(E)$), but there must be *some* $\mu = \mu_0$ which meets E but *not* H. Otherwise the μ's would establish a homeomorphism between E and H, which is impossible since E is compact and non-empty, while H is non-compact or empty. Choose $\gamma = \mu_0 \cap I^+[E]$. Then $\gamma \subset$ int $D^+(E)$ and is future-inextendible as required.

3. THE THEOREM

We shall begin by giving a precise statement of our theorem. The form of statement we adopt is made primarily for the sake of generality and for certain mathematical advantages. But in order that the theorem may be directly applied to physical situations, we single out the main special cases of interest in a corollary. This recasts our main result in a much more suggestive and immediately usable form. However, the generality of the statement given in the theorem will also yield some advantages as regards applications. It will enable a small amount of information to be extracted as to the actual nature of the space-time singularities. Also, it is by no means impossible that the theorem, as stated, may have relevance in physical situations other than precisely those which we have considered here. We shall follow the statement of the theorem with some explanations and interpretations.

THEOREM. *No space-time M can satisfy all of the following three requirements together:*

(3.1) *M contains no closed timelike curves,*

(3.2) *every inextendible causal geodesic in M contains a pair of conjugate points,*

(3.3) *there exists a future- (or past-) trapped set $S \subset M$.*

Let us examine each of these three conditions in turn. With regard to (3.1), the existence of closed timelike curves in any space-time model leads to very severe interpretative difficulties. It might perhaps be argued that the presence of a closed timelike world-line could be admissable, provided the world-line entered a region of such extreme physical conditions, or involved such large accelerations, that no physical observer could 'survive' making this trip into his own past, so that any 'memory' of events would necessarily be destroyed in the course of the trip. However, it seems highly unlikely that the physical consequences of closed timelike curves can be eliminated by considerations of this kind. The existence of such curves can imply serious global consistency conditions on the solutions of hyperbolic differential equations.† We are reassured by the theorem referred to as II in § 1 (cf. **H**) that the singularity problem of general relativity is not forcing us into consideration of closed timelike curves.

Condition (3.2) of the theorem—namely that for any timelike or null geodesic, there is a 'neighbouring geodesic' which meets it at two distinct points—may, at *first sight* appear to be a strong one. However, this is not so. The condition is in fact one that could be expected to hold in *any* physically realistic non-singular space-time. It is a consequence of three requirements: *causal geodesic completeness*, the *energy condition* and a *generality* assumption.

The requirement of *causal geodesic completeness* is simply that every timelike and null geodesic can be extended to arbitrarily large affine parameter value both into the future and into the past. (In the case of timelike geodesics we can use the proper time as such a parameter.) In crude terms we could interpret this condition as saying: 'photons and freely moving particles cannot just appear or disappear off the edge of the universe'. A completeness condition of this kind is sometimes used as virtually a *definition* of what is meant by a non-singular space-time (cf. Geroch 1968*a*). Since one must normally 'delete' any actual singular points from consideration as part of the space-time manifold, it is by some criterion such as 'incompleteness' that the 'holes' left by the removal of the singularities may be detected.

The *energy condition* may be expressed as

$$t^a t_a = 1 \textit{ implies } R_{ab} t^a t^b \leqslant 0. \tag{3.4}$$

(We use a $+---$ signature, with Riemann and Ricci tensor signs fixed by $2\nabla_{[a}\nabla_{b]}k_c = k_d R^d_{aab}$, $R_{ab} = R^c_{acb}$.) With Einstein's equations

$$R_{ab} - \tfrac{1}{2}Rg_{ab} = -KT_{ab}, \tag{3.5}$$

(3.4) becomes

$$t^a t_a = 1 \textit{ implies } T_{ab} t^a t^b \geqslant \tfrac{1}{2}T^c_c. \tag{3.6}$$

(We have $K > 0$. To incorporate a cosmological constant λ, we would have to replace T_{ab} in the above by $T_{ab} + \lambda K^{-1} g_{ab}$. Thus, (3.6), as it stands, would still imply (3.4) so long as $\lambda \leqslant 0$.) [If, in an eigentetrad of T_{ab}, E denotes the *energy density* and p_1, p_2, p_3 denote the three *principal pressures*, then (3.6) can be written as

---

† For example, ϕ = const. is the only solution of $\partial^2\phi/\partial t^2 - \partial^2\phi/\partial x^2 = 0$, on the (x, t)-torus, for which (t, x) is identified with $(t+n, x+m\pi)$ for each pair of integers n, m.

$$E + \Sigma p_i \geqslant 0, \tag{3.7}$$

together with

$$E + p_i \geqslant 0, \tag{3.8}$$

where $i = 1, 2, 3$.

The *weak* energy condition is

$$l^a l_a = 0 \text{ implies } R_{ab} l^a l^b \leqslant 0, \tag{3.9}$$

which is a consequence of (3.4) (as follows by a limiting argument). This is equivalent, assuming Einstein's equations, to (3.8) (*without* (3.7)) and follows from the *positive-definiteness* of the energy expression $T_{ab}t^a t^b$, for $t^a t_a = 1$. (This is now irrespective of the value of λ.)

The assumption of *generality* we require (compare Hawking 1966*b*) is that every causal geodesic γ contains some point for which

$$k_{[a} R_{b]cd[e} k_{f]} k^c k^d \neq 0, \tag{3.10}$$

where k_a is tangent to γ. If γ is timelike, we can rewrite (3.10) as

$$R_{abcd} k^b k^c \neq 0. \tag{3.11}$$

(To see this, transvect (3.10) with $k^a k^f$.)

In any physically realistic 'generic' model, we would expect (3.10) to hold for each γ. For example, the condition can fail for a timelike geodesic γ *only* if $R_{ab}k^a k^b$ vanishes at every point on γ, and then only if the Weyl tensor is related in a very particular way to γ (i.e. $C_{abcd}k^b k^c = 0$) at *every* point on γ. (For a generic space-time this would not even occur at *any* point of any γ!) The condition can fail for a null geodesic γ only if $R_{ab}k^a k^b$ vanishes at every point of γ and the Weyl tensor has the tangent direction to γ as a principal null direction at *every* point of γ (cf. **P**, p. 162). (In a generic space-time, there would not be *any* null geodesic γ which is directed along a principal null direction at *six* or more of its points. This is because null geodesics form a five-dimensional system. It is n conditions on a null geodesic that it be directed along a principal null direction at n of its points, so such null geodesics form a $(5-n)$-dimensional system in a generic space-time.) We can thus reasonably say that it is only in very 'special' (and therefore physically unrealistic) models that the condition will fail.

We must now show why these three conditions together imply (3.2). The fact that they do is essentially a consequence of the *Raychaudhuri* effect (1955, cf. also **P**, p. 169; compare also Myers 1941). The idea here is to proceed so far along the causal geodesic γ that we get beyond the focal length of the effective 'lens system' due to the curvature along γ (compare Penrose 1965*b*). Consider a causal geodesic γ belonging to a hypersurface orthogonal congruence Γ of causal geodesics. We are interested in the members of Γ only in the immediate neighbourhood of γ.

When γ is a null geodesic, we shall, for convenience, specify that all the other members of Γ shall also be null. In this case we shall, in fact, be interested only in those members of Γ, near γ, which generate a null hypersurface containing γ. When γ is time-like we define the vector field t^a to be the unit future-directed tangents to the curves of Γ. When γ is null, we choose a vector field l^a to be smoothly varying future-directed tangents to the curves of Γ, where l^a is parallelly propagated along each curve. We have

$$\nabla_a t_b = \nabla_b t_a, \quad t^a t_a = 1, \quad Dt^a = 0, \quad \text{with} \quad D = t^a \nabla_a \tag{3.12}$$

and

$$l_{[c} \nabla_a l_{b]} = 0, \qquad l^a l_a = 0, \quad Dl^a = 0, \quad \text{with} \quad D = l^a \nabla_a \tag{3.13}$$

respectively.

Let us first consider the timelike case. Ricci identities give, with (3.12),

$$R_{abcd} t^b t^d = D(\nabla_c t_a) + (\nabla_c t^d)\,(\nabla_d t_a). \tag{3.14}$$

Now $R_{abcd} t^b t^d$ and $\nabla_c t_d$ each annihilate t^a when transvected with it on any free index. Introduce an orthonormal basis frame, with t^a as one of the basis elements. Let $Q_{\alpha\beta}$ and $U_{\alpha\beta}$ denote the symmetric (3×3) matrices of spatial components of $R_{abcd} t^d t^d$ and $\nabla_a t_b$, respectively. Then (3.14) becomes

$$Q_{\alpha\beta} = DU_{\alpha\beta} - U_{\alpha\gamma} U_{\gamma\beta}. \tag{3.15}$$

The matrix $Q_{\alpha\beta}$ defines the geodesic deviation (relative acceleration) of Γ; the trace-free part of $U_{\alpha\beta}$ defines the shear of Γ. We define the *divergence* θ of Γ to be

$$\theta = \nabla_a t^a = -U_{\alpha\alpha}. \tag{3.16}$$

Taking the trace of (3.15), we get

$$D\theta + \tfrac{1}{3}\theta^2 = \tfrac{1}{3}\,(U_{\alpha\beta} U_{\alpha\beta} \delta_{\rho\sigma} \delta_{\rho\sigma} - U_{\alpha\beta} \delta_{\alpha\beta} U_{\rho\sigma} \delta_{\rho\sigma}) - Q_{\gamma\gamma} \leqslant 0 \tag{3.17}$$

by Schwarz's inequality and the energy condition (3.4) (which asserts $Q_{\gamma\gamma} \geqslant 0$). Equality holds only when $Q_{\gamma\gamma} = 0$ *and* $U_{\alpha\beta}$ is proportional to $\delta_{\alpha\beta}$ (so that the shear would have to vanish).

Suppose $R_{abcd} t^b t^d \neq 0$ at some point x of γ, in accordance with (3.11). Then $Q_{\alpha\beta} \neq 0$ at x. We shall show, first, that this implies that the *strict* inequality holds in (3.17) at some point y on γ with $x \prec y$. For if it turns out that $Q_{\alpha\beta} = \mu \delta_{\alpha\beta}$ at x (for some μ), then clearly $Q_{\alpha\beta} \neq 0$ at x implies $Q_{\gamma\gamma} \neq 0$ at x, so that strict inequality holds at $y = x$. On the other hand, suppose $Q_{\alpha\beta}$ is *not* of this form at x. Then by (3.15) $U_{\alpha\beta}$ cannot be proportional to $\delta_{\alpha\beta}$ throughout any open segment of γ whose closure includes x. Thus, the expression in parentheses in (3.17) must fail to vanish at some point $y \in \delta$ with $x \prec y$, so the strict inequality in (3.17) must hold at y.

Let the real quantity W be defined along γ as a non-zero solution of

$$DW = \tfrac{1}{3}\theta W \tag{3.18}$$

(so that W^3 measures a spacelike 3-volume element orthogonal to γ and Lie transported along the curves of Γ). Then (3.17) gives†

$$D^2 W \leqslant 0 \tag{3.19}$$

along γ, provided W remains positive. Furthermore the strict inequality holds at y. Choosing $W > 0$ at x, we see from (3.18) and (3.19) that if $\theta \leqslant 0$ at x, then W becomes zero at some point q on γ with $x \ll q$. Furthermore, if $\theta > 0$ at x, then W becomes zero at some $p \in \gamma$ with $p \ll x$. This is *provided* we assume that γ is a *complete* geodesic. (By (3.12), we can interpret the 'D' in (3.17), (3.18), (3.19) as $\mathrm{d}/\mathrm{d}s$, where s is a proper time parameter on γ. The completeness condition ensures that the range of s is unbounded.) When W becomes zero, we have a *focal point* of Γ (point of the caustic) at which θ becomes infinite (since $\theta = 3D \ln W$).

Now fix the causal geodesic γ and fix a point x on it at which (3.11) holds: then allow the congruence Γ to vary. Thus, we consider solutions of (3.15), where the matrix $Q_{\alpha\beta}$ is a given function of s. We shall be interested, in the first instance, in solutions for which $\theta \geqslant 0$ at x. Then by the above discussion there will be a first focal point q_Γ on γ, for each Γ (with $x \ll q_\Gamma$). Each solution of (3.15) is fixed once the value of $U_{\alpha\beta} = \mathring{U}_{\alpha\beta}$ is fixed at x (with $\mathring{U}_{\alpha\alpha} \geqslant 0$). Thus, q_Γ is a function of the nine $\mathring{U}_{\alpha\beta}$. Furthermore, it must be a continuous function. We note that if any component of $\mathring{U}_{\alpha\beta}$ is very large, then q_Γ is very near x (since, in the limit $Q_{\alpha\beta}$ becomes irrelevant and the solution resembles the flat space-time case). It follows that the q_Γ's must lie in a bounded portion ζ of γ. (The one-point compactification of the space of $\mathring{U}_{\alpha\beta}$, with $\mathring{U}_{\alpha\alpha} \geqslant 0$ is mapped continuously into γ, with the point at infinity being mapped to x itself. Thus, the image must be compact.) Choose a point $q \in \gamma$, to the future of ζ and let Γ consist of the timelike geodesics (near γ) through q. If there were no conjugate point to q on γ, then the Γ congruence would be non-singular to the past of q. We cannot have $\theta \leqslant 0$ at x, since this would imply $q \in \zeta$. But we have seen that $\theta > 0$ implies another focal point to the past of x. This establishes the existence of a pair of conjugate points on γ in the timelike case.

When γ is null, the argument is essentially similar. In place of (3.14) we can use the Sachs equations (cf. **P**, p. 167) which have a matrix form similar to (3.15). The components of the curvature tensor which enter into these equations are just the four independent real (or two independent complex) components of $l_{[a} R_{b]cd[e} l_{f]} l^c l^d$. The analogue of θ is $-2\rho = \nabla_a l^a$. In place of W we have a 'luminosity parameter' L, satisfying $DL = -\rho L$ and $D^2 L \geqslant 0$. The conclusion is the same: If (3.10) holds at some point on γ, if γ is complete and if the energy condition holds (in this case the weak energy condition (3.9) will suffice), then γ contains a pair of conjugate points.

---

† Equation (3.19), which follows from $R_{ab} k^a k^b \leqslant 0$, is essentially the statement that 'gravitation is always attractive' (cf. §1). It tells us that the geodesics of Γ, neighbouring to γ, have a tendency to accelerate towards γ—in the sense that freely falling 3-volumes accelerate inwards.

We now come to (3.3), the final condition of the theorem. A drawback of this condition, when it comes to applications, is that we may require considerable information of a global character concerning the space-time M, in order to decide whether or not a given set S is future-trapped. However, in certain special cases, we can invoke the weak energy condition and null-completeness, to enable us to infer, on the *basis of these two properties*, that a certain set should be future-trapped. An example of such a set S is a *trapped surface* (Penrose 1965*a*; **P**, p. 211), defined as a compact spacelike 2-surface with the property that *both* systems of null geodesics which intersect S orthogonally *converge* at S, as we proceed into the future. (For simplicity, suppose S to be achronal.) We expect trapped surfaces to arise when a gravitational collapse of a localized body (e.g. a star) to within its Schwarzschild radius takes place, which does not deviate too much from spherical symmetry. The significant feature of a trapped surface arises from the fact that the null geodesics meeting it orthogonally are the generators of $E^+(S)$. If these null geodesics start out by converging ($\rho > 0$) then by the earlier discussion (Raychaudhuri effect in the null case—weak energy condition and null completeness assumed), they must continue to converge until they encounter a focal point. Either then, or before then, they must leave $E^+(S)$ (cf. **P**, p. 218). Since S is compact and since the focal points must move continuously with the geodesic (being obtainable via integration of curvature), it follows that the geodesic segments joining S to the focal points must sweep out a compact set. Thus $E^+(S)$, being the intersection of this compact set with the closed set $\dot{I}^+[S]$, must also be compact—so S is future-trapped and the theorem applies.

Precisely the same argument will apply in more general situations. For example, if S is any compact achronal set whose edge is smooth and at which the null geodesics which form the local boundary of its future (these will be orthogonal to edge(S)) *converge* at edge(S) as we proceed into the future, then (again *assuming* null completeness and the weak energy condition) S will be future-trapped. More generally still, we need not require that the null geodesics which form the local boundary of the future of S actually converge *at* edge(S). It is only necessary that we should have some reason for believing that they converge *somewhere* to the future of S. In particular, S might contain but a single point p, located somewhere near the centre of a collapsing body, but at a time before the collapse has drastically affected the geometry at p. Then, under suitable circumstances the future null cone of p can encounter sufficient collapsing matter that it (locally) starts converging again. Thus every null geodesic through p will encounter a point conjugate to p in the future (assuming null completeness and the weak energy condition), so again these null geodesic segments sweep out a compact set. Its intersection with $\dot{I}^+(p)$ is $E^+(\{p\})$, implying that $E^+(\{p\})$ is compact, so $\{p\}$ is future-trapped and the theorem applies.

In its time-reversed form, this last example has relevance to cosmology. If the point p refers to the earth at the present epoch, the null geodesics into the past, through p sweep out a region which can be taken to represent that portion of the universe which is visible to us now. If sufficient matter (or curvature in general) encounters these null geodesics, then the divergence $(-\rho)$ of the geodesics may be expected to change sign somewhere to the past of p. This sign change occurs where an object of given size intercepting the null ray subtends its *maximum solid angle* at p. Thus, the *existence* of such a maximum solid angle for objects in each direction, may be taken as the physical interpretation of this type of past-trapped set $\{p\}$. Again the theorem applies. In an appendix we give an argument to show that the required condition on p seems indeed to be satisfied in our universe.

Another example of a future- (or past-) trapped set is any achronal set which is a *compact spacelike hypersurface*. (If we do not assume that the hypersurface is achronal, we can produce a 'copy' of it which *is* achronal by taking a suitable covering manifold of the entire space-time, cf. **H**. Thus, we actually lose no generality by assuming that S is achronal.) In this case, since edge$(S) = \emptyset$, we have $E^+(S) = S$, so $E^+(S)$ is compact. Hence the theorem applies to 'closed universe' models. It is possible that still other situations of physical interest might arise in which a future- (or past-) trapped set S would be inferred as existing (perhaps on the basis of completeness or energy assumptions).

We are now in a position to state the corollary to our theorem.

COROLLARY. A *space-time M cannot satisfy causal geodesic completeness if, together with Einstein's equations* (3.5), *the following four conditions hold:*

(3.20) *M contains no closed timelike curves.*

(3.21) *the energy condition* (3.6) *is satisfied at every point,*

(3.22) *the generality condition* (3.10) *is satisfied for every causal geodesic,*

(3.23) *M contains either*

(i) *a trapped surface,*

or (ii) *a point p for which the convergence of all the null geodesics through p changes sign somewhere to the past of p,*

or (iii) *a compact spacelike hypersurface.*

We may interpret failure of the causal geodesic completeness condition in our corollary as virtually a statement that any space-time satisfying (3.20)–(3.23) 'possesses a singularity' (cf. Geroch 1968*a* and our earlier remarks). However, one cannot conclude, on the basis of the corollary, that such a singularity need necessarily be of the 'infinite curvature' type. Although one might infer that in *some sense* a 'maximally extended' space-time satisfying (3.20)–(3.23) should obtain arbitrarily large curvatures, there are, nevertheless, other possibilities to consider (cf. **H**). In fact, very little is known about the nature of the space-time singularities arising in general relativity other than in highly symmetrical situations. For this reason, it is worth pointing out the minor inference that can be made about the nature of these singularities if we revert back to our original statement of the theorem. The implication is, virtually, that a space-time satisfying

(3.20)–(3.23) must contain a causal geodesic which possesses no pair of conjugate points. At a first guess, one might have imagined that causal geodesics entering very large curvature regions would be inclined to possess many pairs of conjugate points. Instead, we see that our theorem implies that *some* causal geodesic 'enters a singularity' (i.e. is compelled to be geodesically incomplete) before any repeated focusing has time to take place.

Proof of the theorem:

Take S as future-trapped. Then, by lemma (2.12), there is a future-inextendible timelike curve $\gamma \subset \operatorname{int} D^+(E^+(S))$. (That strong causality holds for M follows from lemma (2.10).) Define $T = \bar{I}^-[\gamma] \cap E^+(S)$. We shall show that T is past-trapped. (That T is closed and achronal follows at once since $\bar{I}^-[\gamma]$ is closed and $E^+(S)$ is closed and achronal.) Now, since $\gamma \subset D^+(E(S))$, every past-inextendible timelike curve with future end-point on γ must cross $E^+(S)$. More particularly, it must cross T. Also, $\bar{I}^-[T] \subset I^-[\gamma]$. Thus $I^-[T]$ is simply a portion of $I^-[\gamma]$ 'cut off' by T. Examining the boundaries of these sets, we see $\dot{I}^-[T] \subset T \cup \dot{I}^-[\gamma]$. We are interested in $E^-(T) - T$. This is generated by null geodesics $\{\beta\}$ on $\dot{I}^-[T]$ with future end-point on T (at edge(T)). These null geodesics can be continued on $\dot{I}^-[\gamma]$ inextendibly into the future. (For, by lemma (2.6), each point of $\dot{I}^-[\gamma]$ is the past end-point of a null geodesic on $\dot{I}^-[\gamma]$ which continues future-inextendibly unless it meets γ. But it clearly cannot meet γ, since γ is timelike and future-inextendible.) But, by (3.2), every generator β of $\dot{I}^-[T]$ must, when maximally extended, contain a pair of conjugate points p, q, with $p \prec q$, say. By lemma (2.9), p cannot lie on $\dot{I}^-[\gamma]$ (so $p \in I^-[\gamma]$). Thus β must contain a past end-point either at p, or to the future of p. Now T and edge(T) are compact (being closed subsets of the compact set $E^+(S)$). Since β meets edge(T) and since conjugate points vary continuously, (being obtainable as integrals of curvature, cf. Hicks 1964, **H**) we can choose p and q, for each β, so that the segment of the extension of β from p to q sweeps out a compact region. Thus, the segment of the extension of β from p to edge(T) also sweeps out some compact region C of M. We have $E^-(T) = \dot{I}^-[T] \cap (C \cup T)$, showing that $E^-(T)$ is a closed subset of the compact set $\mathrm{C} \cup T$ and is therefore itself compact. Thus, T is past-trapped, as required.

By lemma (2.12) there exists a past-inextendible timelike curve $\alpha \in \operatorname{int} D^-(E^-(T)$. Choose a point $a_0 \in \alpha$. We have $a_0 \in I^-[\gamma]$, so we find $c_0 \in \gamma$ with $a_0 \ll c_0$. Choose the sequence $a_0, a_1, a_2, \ldots, \in \gamma$, receding into the past indefinitely (i.e. with no limit point). Similarly choose $c_0, c_1, c_2, \ldots \in \gamma$ proceeding into the future indefinitely. We have $a_i \ll c_i$ for all i. Now $a_i \in \operatorname{int} D^-(E^-(T))$ and $c_i \in \operatorname{int} D^+(E^+(S))$. Thus by lemma (2.7) $J^+(a_i) \cap J^-[T]$ is compact (with strong causality holding throughout) and so is $J^-(c_i) \cap J^+[S]$. It is easily seen that $J^+(a_i) \cap J^-(c_i)$, is a closed subset of $\{J^-(c_i) \cap J^+[S]\} \cup \{J^+(a_i) \cap J^-[T]\}$ and so is also compact with strong causality holding throughout. Thus, by lemma (2.11) there is a maximal causal geodesic μ_i from a_i to c_i. Now μ_i must meet T, which is compact, at q_i, say. As $i \to \infty$, there will be an accumulation point q in T and an accumulation causal direction at q.

Choose the causal geodesic μ, through T, in this direction, so μ is approached by μ_i. By (3.2), μ contains a pair of conjugate points, u and v, say, with $u \prec v$. Since conjugate points vary continuously, we must have u as a limit point of some $\{u_j\}$ and v as a limit point of some $\{v_j\}$ where u_j and v_i are conjugate points on the maximal extension of μ_j, the $\{\mu_j\}$ being chosen to converge on μ. But $\{a_i\}$ and $\{c_i\}$ cannot accumulate at any point of the segment uv of μ. Hence, for some large enough j, a_j will lie to the past of u_j in μ_j and c_j to the future of v_j on μ_j. This contradicts lemma (2.8) and the maximality of μ_j. The theorem is thus established.

The authors are grateful to C. W. Misner and to R. P. Geroch for valuable discussions.

REFERENCES

Avez, A. 1963 *Inst. Fourier* **105**, 1.

Boyer, R. H. 1964 *Nuovo Cim.* **33**, 345.

Brans, C. & Dicke, R. H. 1961 *Phys. Rev.* **124**, 925.

Carter, B. 1967 Stationary axi-symmetric systems in general relativity (Ph.D. Dissertation, Cambridge University).

Chandrasekhar, S. 1935 *M.N.* **95**, 207.

Chandrasekhar, S. 1964 *Phys. Rev. Lett.* **12**, 114, 437.

Dicke, R. H. 1962 *Phys. Rev.* **125**, 2163.

Geroch, R. P. 1966 *Phys. Rev. Lett.* **17**, 446

Geroch, R P. 1968*a* *Ann. Phys.* **48**, 526.

Geroch, R. P. 1968*b* *J. Math. Phys.* **9**, 1739.

Hawking, S. W. 1966*a* *Proc. Roy. Soc. Lond.* A **294**, 511

Hawking, S. W. 1966*b* *Proc. Roy. Soc. Lond.* A **295**, 490.

Hawking, S. W. 1966*c* Singularities and the Geometry of space-time (Adams Prize Essay, Cambridge University.)

Hawking, S. W. 1967 *Proc. Roy. Soc. Lond.* A **300**, 187.

Hawking, S. W. & Ellis, G. F. R. 1968 *Astrophys J.* **152**, 25.

Hicks, N. J. 1965 *Notes on differential geometry.* Princeton: D. van Nostrand Inc.

Hoyle, F. & Narlikar, J. V. 1963 *Proc. Roy. Soc. Lond.* A **273**, 1.

Kronheimer, E. H. & Penrose, R. 1967 *Proc. Camb. Phil. Soc. Lond.* **63**, 481.

Lifshitz, E. M. & Khalatnikov, I. M. 1963 *Adv. Phys.* **12**, 185.

Lindquist, R. W. & Wheeler, J. A. 1957 *Rev. Mod. Phys.* **29**, 432.

Milnor, J. 1963 *Morse theory.* Princeton University Press, Princeton.

Myers, S. B. 1941 *Duke Math. J.* **8**, 401.

Penrose, R. 1965*a* *Phys. Rev. Lett.* **14**, 57.

Penrose, R. 1965*b* *Revs. Mod. Phys.* **37**, 215.

Penrose, R. 1966 An analysis of the structure of space-time (Adams Prize Essay, Cambridge University).

Penrose, R. 1968 in *Battelle Rencontres, 1967 Lectures in Mathematics and Physics* (Ed. De Witt, C. M. & Wheeler, J. A.) New York: W. A. Benjamin Inc.)

Penrose, R. 1969 in *Contemporary physics: Trieste Symposium* 1968 (paper SMR/63).

Raychaudhuri, A. 1955 *Phys. Rev.* **98**, 1123.

Seifert, H. J. 1967 *Z. Naturforsch.* **22***a*, 1356.

Sexl, R. U. & Urbantke, H. 1967 *Acta Phys. Austriaca* **26**, 339.

Steenrod, N. 1951 *The topology of fibre bundles* (Princeton University Press).

CHAPTER 42

APPENDIX

We wish to show that there is enough matter on the past light-cone of our present location p to imply that the divergence of this cone changes sign somewhere to the past of p. A sufficient condition for this to be so is that there should be (affine) distances R_1 and R_2 such that along every past-directed null geodesic from p,

$$\tfrac{1}{2}\mathrm{K}R_1 \int_{R_1}^{R_2} T_{ab}\, l^a l^b \,\mathrm{d}r > 1. \tag{A 1}$$

(This formula can be obtained by using a variational approach similar to that used in Hawking (1966*a*).) As in (3.5), $\mathrm{K} = 8\pi G$, where $G\ (= 7.41\times 10^{-29}\ \mathrm{cm\ g^{-1}})$ is the gravitational constant. (Length and time units are related via $c = 1$, i.e. $3\times 10^{10}\ \mathrm{cm} = 1\ \mathrm{s}$.)

In this integral, the vector l^a is a future-directed tangent to the null geodesic and r is a corresponding affine parameter ($l^a \nabla_a r = -1$). Here l^a is parallelly propagated along the null geodesic and is such that $r = 0$ at p and $l^a U_a = 1$, where U^a is the future-directed unit timelike vector representing the local standard of rest at p.

In a recent paper (Hawking & Ellis 1968) it was shown that, with certain assumptions, observations of the microwave background radiation indicate that not only do the past directed null geodesics from us start 'converging again' but so also do the timelike ones. As we are concerned only with the null geodesics, the assumptions we shall need will be weaker.

The observations show that between the wavelengths of 20 cm and 2 mm the background radiation is isotropic to within 1 % and has a spectrum close to that of a black body at 2.7 K. We shall assume that this spectrum and its isotropy indicate not that the radiation was necessarily created with this form, but that it has undergone repeated scattering. (We do not assume that the radiation is necessarily primeval.) Thus there must be sufficient matter on each past directed null geodesic from p to make the optical depth large in that direction. We shall show that this matter will be sufficient to cause the inequality (A 1) to be satisfied.

The smallest ratio of density to opacity at these wavelengths will be obtained if the matter consists of ionised hydrogen in which case there would be scattering by free electrons. The optical depth to distance R would be

$$\int_0^R \frac{\sigma}{m} \rho\, l^a \mathrm{V}_a \,\mathrm{d}r,$$

where σ is the Thomson scattering cross-section, m the mass of a hydrogen atom, ρ the density, measured in g cm^{-3}, of the ionised gas and V^a the local velocity of the gas. The red-shift Z of the gas is given by $(l^a \mathrm{V}_a - 1)$. We assume that this increases down our past-light cone. As galaxies are observed with red-shifts of 0.46 most of the scattering must occur at red-shifts greater than this (in fact if the quasars really are at cosmological distances, the scattering must occur at red-shifts of greater than 2). With a Hubble constant of 100 km s^{-1} Mpc^{-1}, a red-shift of 0.4 corresponds to a distance of about 3×10^{27} cm. Taking R_1 to be this distance, the contribution of the gas density to the integral in (A1) is

$$2.6 \int_{R_1}^{R_2} \rho\, (l^a \mathrm{V}_a)^2\, \mathrm{d}r$$

while the optical depth of gas at red-shifts greater than 0.4 is

$$0.4 \int_{R_1}^{R_2} \rho\, l^a \mathrm{V}_a\, \mathrm{d}r.$$

As $l^a\mathrm{V}_a$ will be greater than 1.4 for $r > R_1$ it can be seen that the inequality (A1) will be satisfied at an optical depth of about 0.1. If the optical depth of the Universe were less than this, one would not expect either a black body spectrum or a high degree of isotropy, as the photons would not suffer sufficient collisions. Even if the radiation arose from an isotropic distribution of black-body emitters at a higher temperature but covering less than $\frac{1}{10}$ of the sky, what one would see would then be a dilute 'grey' body spectrum which could agree with the observations between 20 and 2 cm but which would not fit those at 9 and 2 mm. Thus we can be fairly certain that the required condition is satisfied in the observed Universe.

43

ON QUADRATIC FIRST INTEGRALS OF THE GEODESIC EQUATIONS FOR TYPE {22} SPACETIMES

Comm. Math. Phys. **18(4)** (1970), 265–74. With M. Walker

Here we examine the notions of Killing tensors and Killing spinors so as to illuminate Carter's discovery of a quadratic constant along geodesics in the Kerr space-time and we extend Carter's result, for null geodesics (incorporating the propagation of polarization planes), to other type {22} (or 'type D') metrics. This paper is generalized and simplified in Chapter 54, where some misprints and minor errors in the present paper are corrected on its p. [304]. Perhaps, someday, observations of black holes will become so refined that Carter's constant, or the one we obtain here from the Killing spinor that governs the propagation of photon polarization, can become an observational test of the Kerr metric.

290 On quadratic first integrals of the geodesic equations for type {22} spacetimes. *Comm. Math. Phys.* **18(4)** (1970), 265–74

On Quadratic First Integrals of the Geodesic Equations for Type {*22*} Spacetimes

MARTIN WALKER and ROGER PENROSE
Department of Mathematics, Birkbeck College, London, England

Received May 1, 1970

Abstract. It is shown that every type {*22*} vacuum solution of Einstein's equations admits a quadratic first integral of the null geodesic equations (conformal Killing tensor of valence 2), which is independent of the metric and of any Killing vectors arising from symmetries. In particular, the charged Kerr solution (with or without cosmological constant) is shown to admit a Killing tensor of valence 2. The Killing tensor, together with the metric and the two Killing vectors, provides a method of explicitly integrating the geodesics of the (charged) Kerr solution, thus shedding some light on a result due to Carter.

In a remarkable paper, Carter [1] has shown how to integrate the geodesic equations in a class of solutions of Einstein's equations including the (charged) Kerr solution [2], thus reducing the geodesic problem to one of quadratures. Carter's method depends on a peculiar feature of these solutions, namely that "the Hamilton-Jacobi equation (for the geodesic problem) can be solved by separation of variables", in a particular coordinate system [3]. In the present work we obtain an alternative procedure which achieves the same ends, but which is in some respects more transparent than Carter's. In addition, our method provides an explicit integration of all *null* geodesics in *any* type {*22*} vacuum solution. As yet, however, we have not been able to adapt our method so as to obtain charged particle orbits.

By a *(vacuum)* solution we understand a pair (M, g_{ab}) consisting of a 4-dimensional connected Hausdorff differentiable manifold M, and a pseudo-Riemannian metric g_{ab} of signature $(+ - - -)$ defined on M satisfying Einstein's (vacuum) gravitational field equations (with or without cosmological constant). In order to interpret the physical significance of a vacuum) solution, it is frequently useful to understand the global nature[1] of the spacetime (M, g_{ab}). The global analysis of a (vacuum) solution, however, is usually a difficult task since in most cases g_{ab} is given only locally. A knowledge of four first integrals of the geodesic equations $t^c \nabla_c t^a = 0$ greatly facilitates the extension of a locally given spacetime to a maximal inextendible spacetime (and hence the global analysis of (M, g_{ab})), since in this case an explicit construction can, in principle, be carried out [6]. Theorems which assert the existence of first integrals are, therefore, of interest in general relativity.

---

[1] For example, the nature of singularities [4] and conformal infinity [5].

The spacetime (M, g_{ab}) is said to admit a *first integral of degree m* if a scalar χ, satisfying $t^c \nabla_c \chi = 0$, is defined along any geodesic with parallely propagated tangent vector t^a, where χ is of the form $\chi = K_{a_1 \dots a_m} t^{a_1} \dots t^{a_m}$ for some tensor $K_{a_1 \dots a_m}$. The requirement that χ be constant along geodesics is equivalent to the equation[2]

$$\nabla_{(a} K_{a_1 \dots a_m)} = 0 \,. \tag{1}$$

We shall call a (totally symmetric) solution of Eq. (1) a *Killing tensor of valence m*. When $m = 2$, we shall refer to χ and $K_{a_1 \dots a_m}$ as *quadratic*.

The case $m = 1$ is well known, and Eq. (1) is then Killing's equation $\nabla_{(a} \xi_{b)} = 0$. Moreover, the vector ξ^a has a direct geometrical interpretation as the tangent to the orbit of a point of M under the action of a one-parameter group of isometries of $(M, g_{ab}) : \mathscr{L}_\xi g_{ab} = 0$. No such direct geometrical interpretation of $K_{a_1 \dots a_m}$ for $m > 1$ satisfying Eq. (1) appears to be known.

In certain cases of global analysis, it is sufficient to know the conformal structure of (M, g_{ab}), for example, in the study of conformal infinity [5]. In this case, a knowledge of four first integrals of the null geodesics will suffice. The spacetime (M, g_{ab}) is said to admit a *conformal Killing tensor of valence m* if there exist on M tensors $K_{a_1 \dots a_m}$ and $k_{a_1 \dots a_{m-1}}$ such that

$$\nabla_{(a} K_{a_1 \dots a_m)} = k_{(a a_1 \dots a_{m-2}} g_{a_{m-1} a_m)} \,. \tag{2}$$

Eq. (2) is equivalent to the requirement that $K_{a_1 \dots a_m} l^{a_1} \dots l^{a_m}$ be constant along null geodesics with parallely propagated tangent vector l^a. In particular, for $m = 1$ we have $\nabla_{(a} \xi_{b)} = \varphi g_{ab}$ for some scalar φ, and the conformal Killing vector ξ^a defines a conformal symmetry of (M, g_{ab}) in much the same way that a Killing vector defines an isometry of (M, g_{ab}). Again however, no such interpretation of $K_{a_1 \dots a_m}$ for $m > 1$ satisfying Eq. (2) appears to be known.

It will be observed that any fixed linear combination of symmetrized outer products of (conformal) Killing tensors is again a (conformal) Killing tensor, and that the metric itself is a Killing tensor. A (conformal) Killing tensor which is constructible from the metric and other (conformal) Killing tensors in this way will be called *reducible*, otherwise *irreducible*.

---

[2] Round brackets denote symmetrization of the indices enclosed; square brackets denote antisymmetrization.

In this paper we investigate the case $m = 2$, the existence of quadratic (conformal) Killing tensors, for vacuum solutions whose Weyl tensor[3] is of type *{22}*. We will prove the following:

Theorem 1. *Every type {22} vacuum solution (with or without cosmological constant) admits a quadratic conformal Killing tensor which is irreducible provided the spacetime admits fewer than four independent Killing vectors.*

Since every type *{22}* vacuum solution admits either two or four independent Killing vectors, it is a consequence of Theorem 1 that four first integrals of the null geodesics equations for these solutions can be written down explicitly. More particularly, we also prove

Theorem 2. *The charged Kerr solution with or without cosmological constant* [1, 2, 10] *admits a quadratic Killing tensor which is irreducible provided the angular momentum parameter a is not zero.*

The proofs consist of explicitly constructing the tensor whose existence is asserted. The construction follows from the spinor form of the second Bianchi identity.

Let C_{abcd} be the Weyl tensor of a vacuum solution. Then since $R_{ab} = R^c{}_{acb} = \Lambda g_{ab}$, where R_{abcd} is the Riemann tensor and Λ the cosmological constant, the second Bianchi identity $\nabla_{[a}R_{bc]de} = 0$ becomes $\nabla^a C_{abcd} = 0$. The Weyl spinor $\Psi_{ABCD} = \Psi_{(ABCD)}$ is defined by[4]

$$C_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\overline{\Psi}_{A'B'C'D'}$$

so the spinor Bianchi identities are

$$\nabla^{AA'}\Psi_{ABCD} = 0 \tag{3}$$

where $\varepsilon_{AB} = -\varepsilon_{BA}$, $\varepsilon_A{}^A = 2$, and $\nabla_c\varepsilon_{AB} = 0$ with $\varepsilon_{AB}\varepsilon_{A'B'} = g_{ab}$. Let o^A, ι^A be a spinor dyad satisfying $o_A\iota^A = 1$. Then the real null vectors $l^a = o^A\bar{o}^{A'}$ and $n^a = \iota^A\bar{\iota}^{A'}$ together with the complex null vectors $m^a = o^A\bar{\iota}^{A'}$ and $\bar{m}^a = \iota^A\bar{o}^{A'}$ form a null tetrad $\{l^a, n^a, m^a, \bar{m}^a\}$ on M [12].

The Weyl spinor is completely determined by the five complex scalars $\Psi_0, \ldots, \Psi_4$ of Newman and Penrose [12] defined in terms of Ψ_{ABCD} and the spinor dyad o^A, ι^A. In the case of a type *{22}* Weyl spinor,

---

[3] See, for example, Pirani [9], Penrose [11].

[4] We use "Batelle conventions" for tensors and spinors: see Penrose [11]. Accordingly, latin indices are just *abstract labels* which do not take on numerical values; hence, they do *not* refer to any coordinate system or tetrad basis. On the other hand, Gothic indices do refer to tensor or spinor components in the normal way. These conventions allow one to write the tensor-spinor correspondence without "σ's", e.g., $g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'}$, $\delta_a^b = \varepsilon_A{}^B\varepsilon_{A'}{}^{B'}$, $g^{ab} = \varepsilon^{AB}\varepsilon^{A'B'}$. The spinor "Kronecker delta" is written $\varepsilon_A{}^B$ to avoid possible confusion in raising and lowering its spinor indices.

$\Psi_0 = \Psi_1 = \Psi_3 = \Psi_4 = 0$ provided o^A and ι^A are chosen to be repeated principal spinors of Ψ_{ABCD} (so that the corresponding null vectors l^a and n^a lie along the repeated principal null directions of C_{abcd}). The only freedom, apart from $o^A \to \iota^A$, $\iota^A \to -o^A$, in the choice of o^A and ι^A for a type *{22}* vacuum solution is the two parameter connected subgroup of the Lorentz group acting in the tangent space at each point which leaves the corresponding null directions invariant. The action of this subgroup on o^A, ι^A is given by $o^A \to \lambda o^A$, $\iota^A \to \lambda^{-1}\iota^A$ where λ is a complex scalar. The normalization $o_A \iota^A = 1$ is preserved.

If we let $\psi = 6\Psi_2$, then the Weyl spinor of a type *{22}* vacuum solution is given by

$$\Psi_{ABCD} = \psi o_{(A} o_B \iota_C \iota_{D)} \,. \tag{4}$$

We proceed to prove Theorems 1 and 2 by a sequence of lemmas.

Lemma 1. *Let* $\chi_{BC} = \psi^{-1/3} o_{(B}\iota_{C)}$. *Then*

$$\nabla^{A'}_{(A}\chi_{BC)} = 0 \,. \tag{5}$$

Proof. Consider the quantity $\nabla^{A'}_{(A}[\psi^n o_B \iota_{C)}]$. We shall show that, if $n = -1/3$, every dyad component of this quantity vanishes by virtue of the vanishing of the corresponding component of Eq. (3).

Writing out the Bianchi identity (3) using equation (4) and transvecting with $\psi^{-1}\iota^B o^C o^D$ gives

$$-\tfrac{1}{3}\psi^{-1} o_A \nabla^{AA'}\psi - \tfrac{1}{3}\nabla^{AA'} o_A + \iota^B o^C o^D \cdot [\iota_A \nabla^{AA'}(\iota_{(B} o_C o_{D)}) + o_A \nabla^{AA'}(\iota_{(B}\iota_C o_{D)})] = 0 \,.$$

Expanding out the terms in square brackets and making use of the relation $\varepsilon_A{}^D = \iota^D o_A - o^D \iota_A$, we find

$$\psi^{-1} o_A \nabla^{AA'}\psi - 3 o^D \iota_A \nabla^{AA'} o_D = 0 \,. \tag{6}$$

The analogous component of $\nabla^{A'}_{(A}[\psi^n o_B \iota_{C)}]$ is

$$o^A o^B \iota^C \nabla^{A'}_{(A}[\psi^n o_B \iota_{C)}] = \tfrac{1}{3} n \psi^n [\psi^{-1} o_A \nabla^{AA'}\psi + \tfrac{1}{n} o^D \iota_A \nabla^{AA'} o_D]$$

which vanishes, by Eq. (6), if $n = -1/3$. Therefore

$$o^A o^B \iota^C \nabla^{A'}_{(A}[\psi^{-1/3} o_B \iota_{C)}] = 0 \,,$$

and similarly

$$o^A \iota^B \iota^C \nabla^{A'}_{(A}[\psi^{-1/3} o_B \iota_{C)}] = 0 \,.$$

The components $o^A o^B o^C \nabla^{A'}_{(A}[\psi^{-1/3} o_B \iota_{C)}]$ and $\iota^A \iota^B \iota^C \nabla^{A'}_{(A}[\psi^{-1/3} o_B \iota_{C)}]$ vanish by virtue of the shear-free conditions on o^A and ι^A, which again are consequences of (3). Hence $\nabla^{A'}_{(A}[\psi^{-1/3} o_B \iota_{C)}] = 0$.

Before completing the proof of Theorem 1 we need to translate Eq. (5) into a tensor form. To this end, we define a real symmetric trace-free tensor $P_{ab} = \chi_{AB}\bar{\chi}_{A'B'}$.

Lemma 2. *There exists a vector P^a such that*

$$\nabla_{(a}P_{bc)} = P_{(a}g_{bc)}\,. \tag{7}$$

Proof. We first obtain the identity

$$\tfrac{1}{2}\delta^{(a}_{(a_0}g^{bc)}g_{b_0c_0)} = \delta^{(a}_{a_0}\delta^{b}_{b_0}\delta^{c)}_{c_0} - \varepsilon_{A_0}{}^{(A}\varepsilon_{B_0}{}^{B}\varepsilon_{C_0}{}^{C)}\varepsilon_{A'_0}{}^{(A'}\varepsilon_{B'_0}{}^{B'}\varepsilon_{C'_0}{}^{C')}\,. \tag{8}$$

Denoting the left-hand side of (8) by $E^{abc}_{a_0b_0c_0}$ and the right-hand side by $F^{abc}_{a_0b_0c_0}$, we observe that each is idempotent:

$$E^{abc}_{a_0b_0c_0}E^{a_0b_0c_0}_{a_1b_1c_1} = E^{abc}_{a_1b_1c_1} \qquad F^{abc}_{a_0b_0c_0}F^{a_0b_0c_0}_{a_1b_1c_1} = F^{abc}_{a_1b_1c_1}\,;$$

that

$$E^{abc}_{a_0b_0c_0}F^{a_0b_0c_0}_{a_1b_1c_1} = E^{abc}_{a_1b_1c_1} = F^{abc}_{a_0b_0c_0}E^{a_0b_0c_0}_{a_1b_1c_1}\,;$$

and finally that the traces are equal:

$$E^{abc}_{abc} = F^{abc}_{abc} = 4$$

(since $\delta^{(\alpha}_{\alpha}\delta^{\beta}_{\beta}\delta^{\gamma)}_{\gamma} = \frac{1}{6}n(n+1)(n+2)$ if δ^{β}_{α} is an n-dimensional Kronecker delta). The required equality $E^{abc}_{a_0b_0c_0} = F^{abc}_{a_0b_0c_0}$ now follows. For, the above relations imply that the left range of $F^{abc}_{a_0b_0c_0}$ contains the left range of $E^{abc}_{a_0b_0c_0}$, but also that these ranges have the same dimension (namely 4), so they are equal. Similarly the right ranges are equal. An idempotent is uniquely determined once its left and right ranges are both fixed. Thus Eq. (8) is established.

Now we apply Eq. (8) by transvecting it on both sides with $\nabla_a P_{bc}$. Noting that $\nabla_a P_{bc} = \bar{\chi}_{B'C'}\nabla_{A'A}\chi_{BC} + \chi_{BC}\nabla_{AA'}\bar{\chi}_{B'C'}$ and using (5), we see that $P_{(a_0}g_{b_0c_0)} = \nabla_{(a_0}P_{b_0c_0)}$ where $P_a = \frac{1}{2}g^{bc}\nabla_{(a}P_{bc)}$, which establishes Lemma 2.

Lemma 3. *The conformal Killing tensor P_{bc} is irreducible provided the spacetime admits fewer than four independent Killing vectors.*

Proof. Every type *{22}* solution admits either a two or a four parameter group of isometries [16]. Since the geodesic equations admit four first integrals in the latter case, we restrict our attention to those solutions possessing only two independent Killing vectors. We wish to show that the tensor P_{ab} is not expressible as a linear combination of the metric and the Killing tensors which are symmetric outer products of the two independent Killing vectors. The two Killing vectors span a 2-plane T at each point and, as T is non-null almost everywhere, we need only consider regions in which T is timelike or spacelike.

If T is timelike in some region of the spacetime, then except perhaps on certain (at most) three-dimensional subsets, T does not contain a repeated principal null vector. At some generic point p, let the real null vectors in T be $\alpha^A\bar{\alpha}^{A'}$ and $\beta^A\bar{\beta}^{A'}$. Then, at p, the reducible Killing tensors are linear combinations of $\varepsilon_{AB}\varepsilon_{A'B'}$, $\alpha_A\alpha_B\bar{\alpha}_{A'}\bar{\alpha}_{B'}$, $\alpha_{(A}\beta_{B)}\bar{\alpha}_{(A'}\bar{\beta}_{B')}$, and $\beta_A\beta_B\bar{\beta}_{A'}\bar{\beta}_{B'}$. These three expressions are annihilated when transvected with $\alpha^A\alpha^B\bar{\beta}^{A'}\bar{\beta}^{B'}$, but $\chi_{AB}\bar{\chi}_{A'B'}$ is not since none of $\alpha^A o_A$, $\alpha^A \iota_A$, $\beta^A o_A$, $\beta^A o_A$, or $\beta^A \iota_A$ vanishes. Thus, $\chi_{AB}\bar{\chi}_{A'B'}$ is irreducible when one of the Killing vectors is timelike.

If, on the other hand, T is spacelike in some region, then T will be spanned by complex linear combinations of a pair of complex null vectors $\alpha^A\bar{\beta}^{B'}$ and $\beta^A\bar{\alpha}^{A'}$. It is known in this case that, at a generic point, neither of the repeated principal null vectors l^a or n^a is orthogonal to both spacelike Killing vectors[5]. Therefore, none of the scalars $\alpha^A o_A$, $\alpha^A \iota_A$, $\beta^A o_A$, $\beta^A \iota_A$ vanishes, except perhaps on certain three dimensional subsets of the spacetime. The reducible Killing tensors in this case are linear combinations of $\alpha_A\alpha_B\bar{\beta}_{A'}\bar{\beta}_{B'}$, $\alpha_{(A}\beta_{B)}\bar{\alpha}_{(A'}\bar{\beta}_{B')}$, $\beta_A\beta_B\bar{\beta}_{A'}\bar{\beta}_{B'}$, and $\varepsilon_{AB}\varepsilon_{A'B'}$, and the fact that $\chi_{AB}\bar{\chi}_{A'B'}$ is irreducible then follows, as before, by transvection with $\alpha^A\alpha^B\bar{\alpha}^{A'}\bar{\alpha}^{B'}$ at a generic point.

This completes the proof of Theorem 1.

A quantity such as χ_{BC}, satisfying (5), could be called a *Killing spinor* and has some interest in its own right. Note that if

$$\kappa^A\bar{\kappa}^{A'}\,\nabla_{AA'}\kappa_B = 0\,, \tag{9}$$

then

$$\kappa^A\bar{\kappa}^{A'}\,\nabla_{AA'}(\kappa^B\kappa^C\chi_{BC}) = 0 \tag{10}$$

by virtue of Eq. (5). If Eq. (9) holds, then the null vectors $k^a = \kappa^A\bar{\kappa}^{A'}$ are tangent to null geodesics γ (since (9) implies $k^c\,\nabla_c k^a = 0$). Eq. (10) tells us that the complex scalar $\chi = \kappa^B\kappa^C\chi_{BC}$ is constant along γ. The real scalar $|\chi|^2$ is the conformal quadratic first integral obtained from the conformal Killing tensor P_{ab}. But in addition, $\arg\chi$ is also constant along γ and this gives us further information. We have seen that the constancy of $|\chi|^2$ helps us to locate the null geodesics in M. In an analogous way, the constancy of $\arg\chi$ enables us to determine the parallel propagation of "flag planes" (i.e., polarization planes) along γ. For, Eq. (9) states, in addition to the fact that the k^a are tangent to null geodesics, that the flag plane (cf., [9, 11]) of κ^A is also parallelly propagated along γ. Knowledge of χ_{AB} defines for us this parallel propagation since (for example) if $\chi > 0$ at one point of γ, then $\chi > 0$ also at every other point of γ. This fixes the flag plane of κ^A in relation to χ_{AB}.

---

[5] We wish to thank W. Kinnersley for discussions on this point.

Thus, knowledge of any (non-zero) χ_{AB} satisfying (5) supplies the information of the propagation of polarization of (say) a photon along any null geodesic γ.

The same would also hold for any "Killing spinor" with arbitrary (unequal) numbers of (symmetric) upper and lower indices:

$$\nabla^{(A'_0}_{(B_0}\chi^{A'_1\dots A'_r)}_{B_1\dots B_s)}=0\,, \tag{11}$$

with

$$\chi=\kappa^{B_1}\dots\kappa^{B_s}\bar{\kappa}_{A'_1}\dots\bar{\kappa}_{A'_r}\chi^{A'_1\dots A'_r}_{B_1\dots B_s}\,.$$

Eq. (11) is of interest also in other contexts. In flat spacetimes it represents in a conformally invariant way the general trace-free symmetric *twistor* [13, 14] of valence $\binom{r}{s}$.

Eq. (7) can be put in a form which is more suitable for the calculations involved in the proof of Theorem 2 as follows. Letting

$$o^A\bar{o}^{A'}=l^a,\qquad \iota^A\bar{\iota}^{A'}=n^a,\quad\text{and}\quad 6^{-2/3}\Psi_2{}^{-1/3}\bar{\Psi}_2{}^{-1/3}=\psi^{-1/3}\bar{\psi}^{-1/3}=\Sigma$$

we have

$$P_{bc}=\Sigma(l_{(b}n_{c)}-\tfrac{1}{4}g_{bc})\,, \tag{12}$$

since P_{bc} is tracefree and $l_an^a=1$. Define $Q_{bc}=\Sigma\, l_{(b}n_{c)}$. Then the equation $\nabla_{(a}P_{bc)}=P_{(a}g_{bc)}$ implies

$$\nabla_{(a}Q_{bc)}=Q_{(a}g_{bc)} \tag{13}$$

where $Q_a=P_a+\frac{1}{4}\nabla_a\Sigma$. Suppose now that there exists a scalar Q such that $Q_a=\nabla_aQ$, and let $K_{bc}=Q_{bc}-Qg_{bc}$; then $\nabla_{(a}K_{bc)}=\nabla_{(a}Q_{bc)}-Q_{(a}g_{bc)}=0$ and K_{bc} is a quadratic Killing tensor.

Lemma 4. *The Kerr solution* [10] *admits K_{bc} as a quadratic Killing tensor.*

Proof. The proof consists in finding the vector Q^a given above and showing that Q_a is a gradient. The Kerr metric has been expressed in a variety of coordinate systems [10, 15, 3]. For the purposes of the explicit calculation to follow, the most convenient of these is

$$g_{ab}=\begin{bmatrix}1-2mrR^{-2} & -1 & 0 & 2mra^2R^{-2}\sin^2\theta\\ -1 & 0 & 0 & a\sin^2\theta\\ 0 & 0 & -R^2 & 0\\ 2mra^2R^{-2}\sin^2\theta & a\sin^2\theta & 0 & -(r^2+a^2+2mra^2R^{-2}\sin^2\theta)\sin^2\theta\end{bmatrix} \tag{14}$$

where $a, b=0, 1, 2, 3$; $x^0=u$, $x^1=r$, $x^2=\theta$, $x^3=\varphi$, and $R^2=r^2+a^2\cos^2\theta$. There are two Killing vectors with components, in these coordinates $\xi^a = (1, 0, 0, 0)$ and $\eta^a = (0, 0, 0, 1)$, corresponding to the fact that g_{ab} is independent of u and φ. The contravariant components of g_{ab} are

$$g^{ab} = \begin{bmatrix} -a^2R^{-2}\sin^2\theta & -(r^2+a^2)R^{-2} & 0 & -aR^{-2} \\ -(r^2+a^2)R^{-2} & (2mr-r^2-a^2)R^{-2} & 0 & -aR^{-2} \\ 0 & 0 & -R^{-2} & 0 \\ -aR^{-2} & -aR^{-2} & 0 & -R^{-2}\sin^{-2}\theta \end{bmatrix}. \tag{15}$$

The repeated principal null vectors l^a and n^a to which the above coordinates $\{u, r, \theta, \varphi\}$ are adapted, have components

$$\begin{aligned} l^a &= (0, -1, 0, 0)\,, \\ n^a &= R^{-2}(r^2+a^2, \tfrac{1}{2}(r^2+a^2-2\cancel{a}mr), 0, a)\,, \end{aligned} \tag{16}$$

and satisfy $l^a \nabla_a l^c = 0$, $l_a n^a = 1$. The latter normalization precludes affine parametrization of n^a. The complex convergence ϱ of the null geodesic congruence defined by l^a is given [12] by $\varrho = -(r - ia\cos\theta)^{-1}$, and the only nonvanishing dyad component of the Weyl spinor is $\Psi_2 = -m\varrho^3$ where m is the mass parameter. Therefore

$$\Psi_2^{-1/3}\bar{\Psi}_2^{-1/3} = m^{-2/3}(r^2 + a^2\cos^2\theta) = m^{-2/3}R^2\,.$$

The scalar Σ of Eq. (12) is given by $\Sigma = (6m)^{-2/3}R^2$. Absorbing the constant factor into Q_{bc} we put $Q_{bc} = \frac{1}{2}R^2 l_{(b}n_{c)}$ so that

$$Q^{bc} = \begin{bmatrix} 0 & r^2+a^2 & 0 & 0 \\ r^2+a^2 & r^2+a^2-2\cancel{a}mr & 0 & a \\ 0 & 0 & 0 & 0 \\ 0 & a & 0 & 0 \end{bmatrix} \tag{17}$$

The calculation of the components of the vector Q^a satisfying $\nabla^{(a}Q^{bc)} = Q^{(a}g^{bc)}$ is greatly simplified by noting[6] that

$$\nabla^{(a}Q^{bc)} = g^{e(a}\partial_e Q^{bc)} - Q^{e(a}\partial_e g^{bc)}$$

in any coordinate system, where ∂_e denotes the ordinary partial derivative $\partial/\partial x^e$.

The vector Q^a is then defined by

$$Q^{(a}g^{bc)} = g^{e(a}\partial_e Q^{bc)} - Q^{e(a}\partial_e g^{bc)}\,.$$

The calculation is further simplified by noting that $Q^{2a} = 0$ so that the $\partial_2 g^{bc} = (\partial/\partial\theta)g^{bc}$ are not required, and also that $\partial_2 Q^{bc} = 0$. We find

$$Q^a = 2rR^{-2}(r^2+a^2, r^2+a^2-2mr, 0, a)$$

---

[6] We wish to thank R. Geroch for pointing out this fact to us.

so that $Q_a = g_{ac}Q^c = (0, -2r, 0, 0) = \nabla_a(-r^2)$. We conclude that the components

$$K^{ab} = Q^{ab} + r^2 g^{ab}$$

define a quadratic Killing tensor for the Kerr solution. Explicitly,

$$K^{ab} = \begin{bmatrix} -r^2a^2R^{-2}\sin^2\theta & (r^2+a^2)R^{-2}a^2\cos^2\theta & 0 & -ar^2R^{-2} \\ (r^2+a^2) \cdot R^{-2}a^2\cos^2\theta & (r^2+a^2-2mr) \cdot R^{-2}a^2\cos^2\theta & 0 & a^3R^{-2}\cos^2\theta \\ 0 & 0 & -r^2R^{-2} & 0 \\ -ar^2R^{-2} & a^3R^{-2}\cos^2\theta & 0 & -r^2R^{-2}\sin^{-2}\theta \end{bmatrix}. \quad (18)$$

In order to complete the proof of Theorem 2 we note that the contravariant components of the metric of the charged Kerr solution with cosmological constant are obtained from those of the metric of the Kerr solution by replacing $2mr - r^2 - a^2$ in g^{11} of Eq. (15) by $2mr - r^2 - a^2 + e^2 + \frac{1}{3}\Lambda r^4$, where e is the charge parameter and Λ the cosmological constant. Replacing K^{11} in Eq. (18) by $R^{-2}(r^2 + a^2 - 2mr - e^2 - \frac{1}{3}\Lambda r^4) \cdot a^2\cos^2\theta$, a straightforward calculation shows that $\nabla^{(a}K^{bc)} = 0$ in the more general case also.

Acknowledgements. We wish to thank P. Havas and R. Geroch for stimulating discussions during the course of this work.

References

1. Carter, B.: Commun. Math. Phys. **10**, 280 (1968).
2. Newman, E. T., Couch, E., Chinnapared, R., Exton, A., Prakash, A., Torrence, R.: J. Math. Phys. **6**, 918 (1965).
3. Carter, B.: Phys. Rev. **174**, 1559 (1968).
4. Geroch, R.: Ann. Phys. **48**, 526 (1968).
5. Penrose, R.: Proc. Roy. Soc. A **284**, 159 (1965).
6. Geroch, R.: J. Math. Phys. **9**, 450 (1968).
7. Eisenhart, L. P.: Riemannian geometry, p. 128. Princeton, N. J.: Princeton University Press 1966.
8. Schouten, J. A.: Ricci calculus, 2nd edition, p. 293. Berlin, Göttingen, Heidelberg: Springer 1954.
9. Pirani, F. A. E.: Introduction to gravitational radiation theory, Lectures on general relativity, Brandeis Summer Institute (1964), Vol. I, S. Deser and K. Ford editors. Englewood Cliffs, N. J.: Prentice-Hall, Inc. 1965.
10. Kerr, R. P.: Phys. Rev. Letters **11**, 237 (1963).
11. Penrose, R.: Structure of space-time. Battelle rencontres. C. de Witt and J. A. Wheeler, Eds. New York: W. A. Benjamin Inc. 1968.
12. Newman, E. T., Penrose, R.: J. Math. Phys. **3**, 566 (1962).

13. Penrose, R.: J. Math. Phys. **8**, 345 (1967).
14. — Int. J. Theor. Phys. **1**, 61 (1968).
15. Boyer, R. H., Lindquist, R. W.: J. Math. Phys. **8**, 265 (1967).
16. Kinnersley, W.: J. Math. Phys. **10**, 1195 (1969).

M. Walker,
R. Penrose
Department of Physics
University of Texas at Austin
Austin, Texas 78712, USA

44

SCATTERING OF TWO IMPULSIVE GRAVITATIONAL PLANE WAVES

Nature **229(5281)** (1971), 185. With K. A. Khan

Scattering of Two Impulsive Gravitational Plane Waves

IT has long been recognized that Einstein's gravitational equations admit wave-like solutions. But because of the smallness of the gravitational constant it would require an exceedingly violent astronomical process to generate gravitational waves of sufficient intensity to be detectable with available apparatus. Weber[10] has, however, presented a plausible case for believing that his detectors may indeed have monitored frequent and sharply pulsed gravitational waves, which seem to be emanating from the centre of our galaxy. Because such waves could give some very significant information about extraordinary processes which may be taking place in certain regions of the universe, the detection of gravitational waves may be important in the observational astronomy of the future.

Unlike the waves of standard electromagnetic theory, however, gravitational waves cannot generally be superposed linearly, except in the weak field limit. A significant theoretical requirement, therefore, is to determine the way in which gravitational waves can scatter one another. But so far, rather little is known about the detailed nature of this scattering. The purpose of this communication is to try to remedy this situation to some degree. We shall present, for the first time, an exact solution of Einstein's vacuum equations representing a collision between two impulsive (δ-function) gravitational plane waves. This facilitates detailed examination of the scattering for this case.

The metric

$$ds^2 = dU\{2dV + (Y^2 - X^2)h(U)dU\} - dX^2 - dY^2 \qquad (1)$$

due originally to Brinkmann[2], and interpreted by Robinson[8], represents a linearly polarized gravitational plane wave. The amplitude function $h(U)$ can be defined arbitrarily, its values giving a direct measure of the curvature of the space–time. Two such waves moving in the same direction can be superposed, simply by adding their corresponding h functions. Thus, when moving in the same direction, two such gravitational waves do not scatter one another.

To exhibit scattering we need waves moving in different directions. If we consider sandwich waves[1], we can match together the regions of flat space which precede each of the two waves. For this, the Rosen[9] form of the metric is more convenient than equation (1). By means of the coordinate change[8] $U=u$, $V=v+\frac{1}{2}x^2FF'+\frac{1}{2}y^2GG'$, $X=xF$, $Y=yG$, where F and G are functions of U subject to $F''=hF$, $G''=-hG$, we obtain Rosen's form

$$ds^2=2L\mathrm{d}u\mathrm{d}v-F^2\mathrm{d}x^2-G^2\mathrm{d}y^2 \qquad (2)$$

(with $L=1$). If $F=1+u\theta(u)$ and $G=1-u\theta(u)$, where $\theta(u)$ is Heaviside's step function, we get the "impulsive wave"[6] for which $h(u)=\delta(u)$. Now, by allowing L, F and G to be functions of both u and v in (2) we can incorporate waves moving in both directions. If L, F and G are functions of u alone for $v<0$, and functions of v alone for $u<0$ (with, say, $L=F=G=1$ for $u<0$, $v<0$) then the metric (2) describes a collision between two waves. If, in particular, we choose each to be an impulsive plane wave, we obtain the following explicit solution of Einstein's vacuum equations[4]

$$\begin{aligned}\mathrm{d}s^2 = \frac{2t^3\mathrm{d}u\mathrm{d}v}{rw(pq+rw)^2} &- t^2\left(\frac{r+q}{r-q}\right)\left(\frac{w+p}{w-p}\right)\mathrm{d}x^2 \\ &- t^2\left(\frac{r-q}{r+q}\right)\left(\frac{w-p}{w+p}\right)\mathrm{d}y^2 \qquad (3)\end{aligned}$$

where

$$p=u\theta(u),\quad q=v\theta(v),\quad r=(1-p^2)^{\frac{1}{2}},\quad w=(1-q^2)^{\frac{1}{2}},$$
$$t=(1-p^2-q^2)^{\frac{1}{2}}=(r^2-q^2)^{\frac{1}{2}}=(w^2-p^2)^{\frac{1}{2}}$$

are C^0 (or C^1) functions which are piecewise C^∞, in the relevant region. The region of validity of the metric (3) is the range "below" $v=1$, $u^2+v^2=1$, $u=1$, as indicated in Fig. 1. (This is the union of the three regions ($v<1$, $u<0$), ($u<1$, $v<0$), ($v^2+u^2<1$).) The curvature for the metric (3) involves δ-functions along $u=0$ and along $v=0$. A singularity near which curvature scalars approach infinity occurs at $u^2+v^2=1$, $u>0$, $v>0$. The regions $v=1$, $u<0$ and $u=1$, $v<0$ are, surprisingly, also singular. For the $u=1$ singularity is actually a region of focusing of the $v=0$ wave by the $u=0$ wave[7]. (In the absence of the $v=0$ wave, but with the same type of coordinates as used in (3), $u=1$ would have been a spurious coordinate singularity.) There is a finite jump in the curvature tensor at $u=0$, $0<v<1$ and at $v=0$, $0<u<1$, in addition to the delta function there. The metric (3) in fact describes the most general space–time representing a collision between parallel-polarized impulsive gravitational waves.

In summary, two gravitational impulsive plane waves approaching each other from different directions will certainly scatter one another. They will cease to be plane after collision and their pure δ-function form becomes modified by the

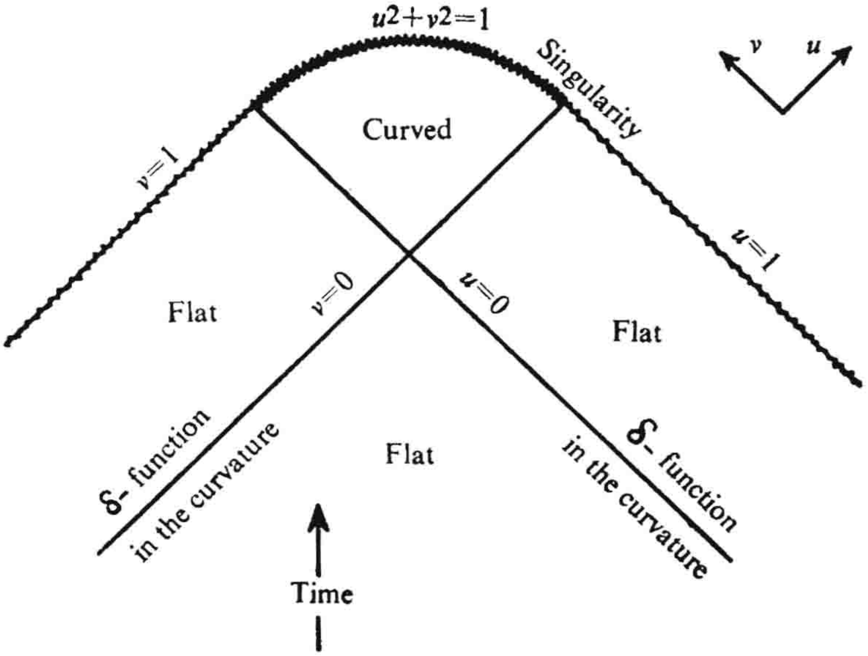


Fig. 1 The (u,v)-plane for colliding impulsive gravitational waves.

addition of a discontinuity in the curvature. Eventually the focusing effect of each wave on the other results in a singularity in space–time.

It should be remembered that we are dealing with exact plane waves here. For waves which are only approximately plane, it is not clear that the situation would in general be qualitatively similar. The singularity which arises here might be a feature merely of the exact focusing due to the high symmetry of the plane waves. On the other hand, an apparently analogous situation previously arose in connexion with singularities occurring in gravitational collapse and cosmology. It was first thought that these singularities were special features due to high symmetry in the exact models used, but later it was shown that this is not the case[3]. Thus it is possible that the presence of singularities is a "generic" feature of space–times resulting from collisions between gravitational waves, provided that the waves are in some appropriate sense sufficiently strong to ensure that "serious" focusing takes place. Even so, there would be no reason to believe that such "generic" singularities would resemble the final singularity of metric (3) at all closely.

There is another respect in which we might have to regard the present discussion as rather unrealistic. This lies in the fact that pure δ-function waves are unlikely to be found in nature, even approximately. In order for energy flow to be kept within reasonable limits, we would have to assume an amplitude function whose time-average is zero[5]. Our discussion, however, should not have to be greatly modified in order to treat such situations. There are general arguments which show that if two plane gravitational waves collide, then, on matter what profile their amplitude functions have, the resulting space–time will still resemble that indicated in Fig. 1 in its essentials. In particular, a space–time singularity will still arise even though the amplitude profiles may be perfectly smooth and the curvature in the plane waves everywhere finite.

Thus the "singularity" of the incoming δ-function profiles is of an inessential character in the sense that it can be "smoothed away", unlike (presumably) the singularity at $u^2+v^2=1$. A plane wave of general profile could be approximated by a series of δ-functions. It would be interesting, therefore, to obtain the generalization of (3) which describes a collision between two successions of impulsive waves, since this would be a little closer to reality than the simple case considered here.

Added in proof: P. Szekeres (*Nature*, **228**, 1183; 1970) has recently obtained a result similar to ours for colliding gravitational plane waves with a particular (non-impulsive) profile.

K. A. KHAN*
R. PENROSE

Birkbeck College,
University of London

Received August 11, 1969; revised July 30, 1970.

*Present address: University of Karachi.

[1] Bondi, H., Pirani, F. A. E., and Robinson, I., *Proc. Roy. Soc.*, **251A**, 519 (1959).
[2] Brinkmann, H. W., *Proc. US Nat. Acad. Sci.*, **9**, 1 (1923).
[3] Hawking, S. W., and Penrose, R., *Proc. Roy. Soc.*, **314A**, 529 (1970).
[4] Khan, K. A., thesis, Univ. London (1969).
[5] Landau, L., and Lifshitz, E. M., *The Classical Theory of Fields* (Addison-Wesley, New York, 1951).
[6] Penrose, R., *Structure of Space-Time*, in *Battelle Rencontres* (edit. by DeWitt, C. M., and Wheeler, J. A.) (Benjamin, New York, 1968).
[7] Penrose, R., *Revs. Mod. Phys.*, **37**, 215 (1965).
[8] Ehlers, J., and Kundt, W., *Exact Solutions of the Gravitational Equations*, in *Gravitation: An Introduction to Current Research* (edit. by Witten, L.) (Wiley, New York, 1962).
[9] Rosen, N., *Phys. Z. Sowjet.*, **12**, 366 (1937).
[10] Weber, J., *Phys. Rev. Lett.*, **22**, 1320 (1969); *ibid.*, **24**, 276 (1970).

45

APPLICATIONS OF NEGATIVE DIMENSIONAL TENSORS

Combinatorial mathematics and its applications (Proc. Conf., Oxford, 1969). Ed. D. J. A. Welsh, pp. 221–44. Academic Press, London, 1971

This somewhat exotic excursion into combinatorics was a natural outcome of my investigations into abstract tensor systems as described—without any diagrams—in my Ph.D. thesis (Chapter 6), where 'negative-dimensional' tensor systems are considered, although rather hidden in the thesis; see Section A1.9a on pp. [58], [59] of that work. The binors of Chapter 31 are of dimension –2, and the present article puts them in a broader context, showing how the diagrammatic and abstract-index notations combine naturally, and would have made my Ph.D. thesis much more comprehensible had I used them! The present article reveals my time-wasting early fascination with the 4-colour problem, which was, at that time (before Appel and Haken, 1976), an unproved conjecture. But I obtained many formulae and diagrammatic curiosities that relate both to the 4-colour problem and to the evaluation of the spin networks of Chapter 31. (Note, in particular, the last paragraph of the main text of the present article, and also the footnote on pp. [243] and [244], but I have still not published the results referred to.)

308 Applications of negative dimensional tensors. *Combinatorial mathematics and its applications* (Proc. Conf., Oxford, 1969). Ed. D. J. A. Welsh, pp. 221–44

Applications of Negative Dimensional Tensors

ROGER PENROSE

Birkbeck College, University of London, England

I wish to describe a theory of "abstract tensor systems" (abbreviated ATS) and indicate some applications. Unfortunately I shall only be able to give a very brief outline of the general theory here.†

I take as my model, the conventional tensor index notation with Einstein's summation convention, which has become so familiar in physics and in what is now referred to as "old fashioned" differential geometry. The elements of an ATS may be denoted by kernel symbols with indices in a way formally identical with the tensor index notation, but now the meanings of the indices are to be quite different. This will enable more general types of object than ordinary tensors to be considered. Some of these (for example, "negative dimensional" tensors) will not be representable in terms of components in the ordinary way.

Each index is to be simply a *label* and does not stand for, say, 1, 2, ..., *n*. Thus an element ξ^a (a "vector") of an ATS is not a set of components, but a single element of a vector space (or module) $\mathscr{T}^a$ over a field (or ring) $\mathscr{T}$. Since I wish to mirror the ordinary index notation and allow expressions such as $\xi^a \xi^b$ or $\xi^a \eta^b - \eta^a \xi^b$, for example, I shall also need an element ξ^b distinct from ξ^a, and so on. Thus we need another vector space (or module) $\mathscr{T}^b$ which will be canonically isomorphic with $\mathscr{T}^a$, etc. etc. Let me define the labelling set $\mathscr{L}$:

$$\mathscr{L} = (a, b, c, \ldots, z, a_0, b_0, \ldots, a_1, \ldots.),$$

supposed infinite. The elements of $\mathscr{L}$ are to be the allowed "abstract indices". We shall then have an infinite class of canonically isomorphic modules

$$\mathscr{T}^a \cong \mathscr{T}^b \cong \ldots\ldots$$

where corresponding elements are denoted by $\xi^a, \xi^b, \ldots.$. Thus $\lambda\xi^a + \mu\eta^a = \zeta^a$ iff $\lambda\xi^b + \mu\eta^b = \zeta^b$, etc. where $\lambda, \mu \in \mathscr{T}$; $\xi^a, \eta^a, \zeta^a \in \mathscr{T}^a$; $\xi^b, \eta^b, \zeta^b \in \mathscr{T}^b$; etc. But in addition we shall need to mirror the other tensor operations (e.g. outer product, contraction) and thus to consider many-indexed quantities which look ($a, b, \dots, d, f, \dots, h$ being all distinct) like

---

† A more extended account is to be found in [4]. However, the theory is there made unnecessarily complicated, because of an inconvenient (but apparently natural) choice having been made in connection with the notation. The point is avoided here by the use of infinitely many canonically isomorphic copies of each vector space or module. See [5] in connection with the approach used here. Also, compare [3] as regards diagrammatic notation.

$$\chi^{ab\dots d}_{f\dots\dots h}.$$

Such quantities will be elements of sets $\mathscr{T}^{ab\dots d}_{f\dots h}$, each of these sets again being a module over $\mathscr{T}$. (The ordering of $a, b, \dots d$ and of $f, \dots, h$ is of significance for each $\chi^{ab\dots d}_{f\dots h}$ but not for the *sets* $\mathscr{T}^{a\dots d}_{f\dots h}$. That is: $\mathscr{T}^{ab\dots d}_{f\dots h} = \mathscr{T}^{db\dots a}_{h\dots f}$, for example, but generally $\chi^{ab\dots d}_{f\dots h} \neq \chi^{db\dots a}_{h\dots f}$).

The entire tensor system $\{\mathscr{T}\}$ will have, as elements, the members of the sets

$$(\mathscr{T}, \mathscr{T}^{a}, \mathscr{T}^{b}, \dots, \mathscr{T}^{a_0}, \dots, \mathscr{T}_{a}, \mathscr{T}_{b}, \dots, \mathscr{T}^{a}_{b}, \dots, \mathscr{T}^{xm}_{p}, \dots).$$

There are four basic operations on $\{\mathscr{T}\}$. These are

addition: $\mathscr{T}^{x\dots z}_{u\dots w} \times \mathscr{T}^{x\dots z}_{u\dots w} \to \mathscr{T}^{x\dots z}_{u\dots w}$

(outer) multiplication: $\mathscr{T}^{a\dots d}_{p\dots r} \times \mathscr{T}^{x\dots z}_{u\dots w} \to \mathscr{T}^{a\dots dx\dots z}_{p\dots ru\dots w}$

contraction, e.g. $\mathscr{C}^{p}_{q}$: $\mathscr{T}^{px\dots z}_{qu\dots w} \to \mathscr{T}^{x\dots z}_{u\dots w}$

index substitution: $\mathscr{T}^{x\dots z}_{u\dots w} \to \mathscr{T}^{f\dots h}_{k\dots m}$.

In the first three of these operations, the differently denoted index letters appearing are all assumed to be different elements of $\mathscr{L}$. In the final operation of index substitution, $x, \dots, z, u, \dots, w$ of $\mathscr{L}$ are all distinct and so are $f, \dots, h, k, \dots, m$. Otherwise they are unrestricted except that $x, \dots, z$ and $f, \dots, h$ are equal in number and that $u, \dots, w$ and $k, \dots, m$ are equal in number.

The axioms are: *First* that addition defines an Abelian group structure for each $\mathscr{T}^{\cdots}_{\cdots}$. *Second*, multiplication is distributive over addition and commutative.† (Non-commutative systems can also be considered, but I shall not concern myself with them here.) *Third*, contraction is distributive over addition and it appropriately commutes with multiplication and with other contractions, e.g.

$$\mathscr{C}^{a}_{b}\,(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot b\cdot\cdot\cdot} + \phi^{\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot\cdot\cdot\cdot b\cdot}) = \mathscr{C}^{a}_{b}(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot b\cdot\cdot\cdot}) + \mathscr{C}^{a}_{b}\,(\phi^{\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot\cdot\cdot\cdot b\cdot}),$$

$$\mathscr{C}^{a}_{b}\,(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot b\cdot\cdot\cdot}\,\psi^{\cdot\cdot\cdot\cdot}_{\cdot\cdot\cdot\cdot\cdot}) = \mathscr{C}^{a}_{b}\,(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot}_{\cdot\cdot b\cdot\cdot\cdot})\,\psi^{\cdot\cdot\cdot\cdot}_{\cdot\cdot\cdot\cdot\cdot},$$

$$\mathscr{C}^{a}_{b}\,(\mathscr{C}^{c}_{d}\,(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot c\cdot\cdot\cdot\cdot}_{\cdot\cdot d\cdot\cdot\cdot\cdot\cdot\cdot b\cdot\cdot})) = \mathscr{C}^{c}_{d}\,(\mathscr{C}^{a}_{b}\,(\chi^{\cdot\cdot\cdot a\cdot\cdot\cdot c\cdot\cdot\cdot\cdot}_{\cdot\cdot d\cdot\cdot\cdot\cdot\cdot\cdot b\cdot\cdot})).$$

---

† This is not really inconsistent with the usual algebraic definition of tensor product, which is formally non-commutative. Here, the *labels* cope with the ordering in a tensor product. Thus, whereas $\xi^{a}\,\eta^{b} = \eta^{b}\,\xi^{a}$, we have $\xi^{a}\,\eta^{b} \neq \eta^{a}\,\xi^{b}$.

Fourth, an index substitution is effected by any permutation of $\mathscr{L}$ and leaves the validity of any formula unaltered, e.g. if

$$\chi^{ab}_{c} = \phi^{a}_{c}\,\eta^{b} + \mathscr{C}^{p}_{q}\,(\theta^{bpa}\,\zeta_{cq}),$$

then

$$\chi^{xn}_{b} = \phi^{x}_{b}\,\eta^{n} + \mathscr{C}^{q}_{a}\,(\theta^{nqx}\,\zeta_{ba}).$$

etc. The "dummy index" notation will be used for contraction. Thus we can write the above

$$\chi^{ab}_{c} = \phi^{a}_{c}\,\eta^{b} + \theta^{bpa}\,\zeta_{cp},$$

for example, so that repeated indices (one upper and one lower) do not contribute to the total "label" $\{{}^{ab}_{c}\}$ of the tensor. These operations and axioms define an *abstract tensor system* or ATS.

Let me remark, in passing, one way of generating new tensor systems from old. This is by means of the device of "clumping indices together". Thus, if we consider a new labelling system $\widetilde{\mathscr{L}} = (\mathscr{A}, \mathscr{C}, \mathscr{E}, \ldots, \mathscr{A}_0, \ldots..)$ where, say, $\mathscr{A} = (a, b), \mathscr{C} = (c, d), \mathscr{E} = (e, f), \ldots, \mathscr{A}_0 = (a_0, b_0), \ldots,$ then we can construct a new ATS $\{\widetilde{\mathscr{T}}\}$ whose elements belong to a subset of $\{\mathscr{T}\}$ namely $\widetilde{\mathscr{T}} = \mathscr{T}, \widetilde{\mathscr{T}}^{\mathscr{A}} = \mathscr{T}^{ab}, \ldots.$ More generally, we can also clump together upper *and* lower indices as a single compound index. (For example $\widetilde{\mathscr{T}}^{\mathscr{A}} = \mathscr{T}^{a}_{\ bc}$, $\mathscr{T}^{\mathscr{D}} = \mathscr{T}^{d}_{\ ef}, \ldots, \widetilde{\mathscr{T}}_{\mathscr{A}} = \mathscr{T}^{bc}_{\ \ a}, \ldots..$ Here, each script letter stands for three elements of $\mathscr{L}$ two of which are in reverse position.) The axioms for the new ATS are satisfied in consequence for those for the old. We can also consider ATS's with indices of differing types (e.g. $\chi^{ab\ \alpha AB}_{\ \ \ \sigma}$) where contractions and index substitutions are not permitted between indices of different types. If we consider different kinds of clumping operations applied simultaneously to an ATS, then each different kind of clumping plays the role of a different type of index in the resulting new ATS.

If desired, we may require of an ATS that it possesses "unit elements". For example, an element $1 \in \mathscr{T}$ such that

$$1\chi^{\cdots\cdot}_{\cdots\cdot} = \chi^{\cdots\cdot}_{\cdots\cdot}$$

for all $\chi^{\cdots\cdot}_{\cdots\cdot} \in \mathscr{T}^{\cdots\cdot}_{\cdots\cdot}$ and an element $\delta^{a}_{b} \in \mathscr{T}^{a}_{b}$ such that

$$\chi^{\cdot\cdot p\cdot\cdot}_{\cdots\cdot\cdot}\,\delta^{q}_{p} = \chi^{\cdot\cdot q\cdot\cdot}_{\cdots\cdot\cdot},\ \chi^{\cdots\cdot}_{\cdot\cdot x\cdot\cdot}\,\delta^{x}_{y} = \chi^{\cdots\cdot}_{\cdot\cdot y\cdot\cdot}$$

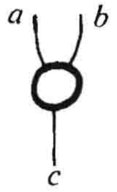
(Thus δ^a_b has the formal properties of a Kronecker delta.) These elements must necessarily be unique. The systems $\{\mathscr{Z}\}$ and $\{\mathscr{I}\}$ generated by δ^b_a and 1, and by δ^b_a and the elements of $\mathscr{T}$, respectively, have a special interest. (These mirror those tensors in an "ordinary" system whose components are absolutely invariant under the full linear group of coordinate transformations, these components being integers in the case of $\{\mathscr{Z}\}$.) We define the "dimension" of $\mathscr{T}^a$ to be the scalar

$$\delta^a_a = \nu.$$

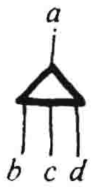
In "reasonable" systems, ν is a non-negative integer. We shall also consider "unreasonable" systems in which ν is a *negative* integer.

A motivation for the above notation is that even in the case of ordinary finite dimensional systems we can retain the full flexibility and simplicity of the tensor index notation while eliminating the undesirable basis dependence of the usual notation. However the above is still subject to the other criticism which is sometimes levelled at such index notations, namely the fact that with many indices, expressions may become cumbersome and all-important index connections are easily misread. I shall therefore introduce a diagrammatic notation for tensors which in most instances allows connections between indices to be discerned at a glance.

Let $a, b, \ldots$ label points in a plane. We denote an object $\chi^{a\ldots c}_{d\ldots f}$ by a symbol with an "arm" corresponding to each of $a, \ldots, c$ and a "leg" corresponding to each of $d, \ldots, f$. For example, if $\theta^{ab}_c \in \mathscr{T}^{ab}_c$, $\chi^a_{bcd} \in \mathscr{T}^a_{bcd}$ we could write

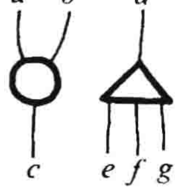


$$= \theta^{ab}_c,$$



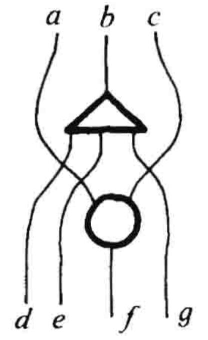
$$= \chi^a_{bcd}.$$

Normally, the label symbols $a, b, c, \ldots$ would be omitted in a diagram. In fact, we may regard the labels simply *as* points of the plane on which the figure is depicted. Outer products are drawn simply as a juxtaposition of the individual symbols:



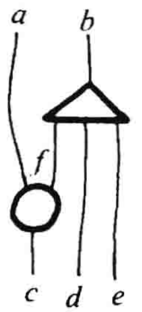
$$= \theta^{ab}_c \chi^d_{efg} \in \mathscr{T}^{abd}_{cefg}$$

or



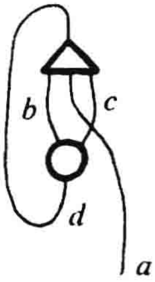
$$=\theta_f^{ac}\,\chi_{deg}^{b}$$

for example. We do not require products to be ordered horizontally. (We should not have had this freedom if we had been concerned with non-commutative ATS's.) To depict a contraction, we simply join the corresponding arm and leg:

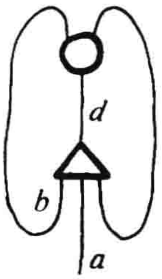


$$=\theta_c^{af}\,\chi_{fde}^{b}\in\mathscr{T}_{cde}^{ab}$$

or



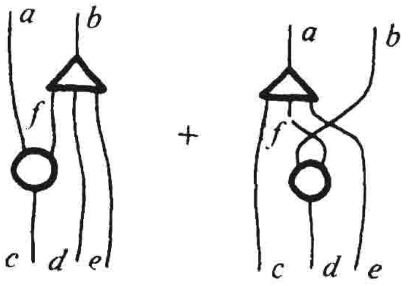
$=$



$$=\theta_d^{bc}\,\chi_{bac}^{d}\in\mathscr{T}_a.$$

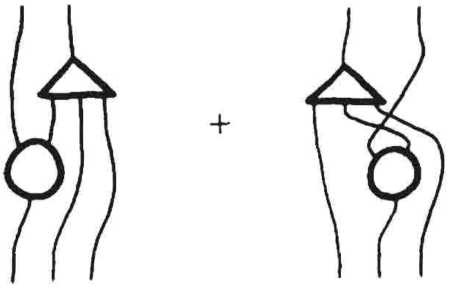
For clarity it will be usual to draw all the free arms (non-dummy upper indices) as emerging at the top of the picture and free legs (non-dummy lower indices) emerging at the bottom. In any case each arm should emerge as finally directed upwards and each leg as finally directed downwards. (As far as possible we try to keep the lines proceeding essentially vertically.) This will cease to apply when we consider Cartesian systems shortly, however.

Normally, to add two expressions it will be convenient simply to draw the diagram for each term and put a "+" sign between them, e.g.

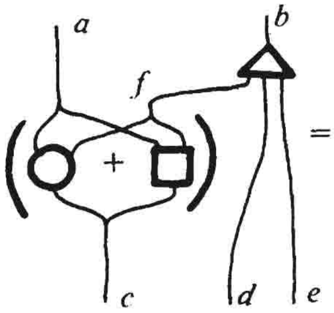


$$= \chi^{b}_{fde}\,\theta^{af}_{c} + \chi^{a}_{cfe}\,\theta^{bf}_{d}$$

We may omit the labels and draw this



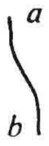
making sure that it is clear, from the arrangements of free arms and legs, which are corresponding to which in the different terms. Occasionally it is convenient to employ a notation



$$= (\theta^{af}_{c} + \gamma^{af}_{c})\,\chi^{b}_{fde}$$

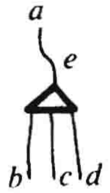
when sums in parenthesis are involved.

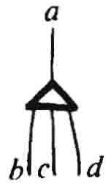
The notation for the unit δ^a_b is simply a "disembodied" line:

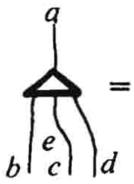


$$= \delta^{a}_{b}$$

since then the rules

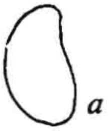
$\chi^e_{bcd}\delta^a_e =$ 

$=$ 

$= \chi^a_{bcd} =$ 

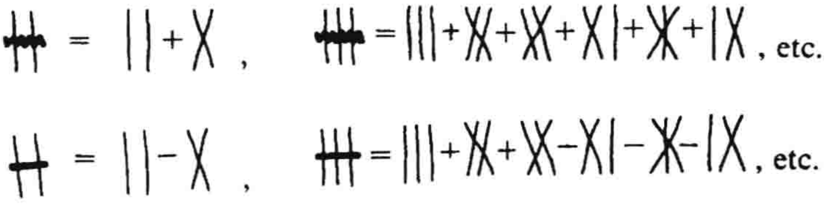
$= \chi^a_{bed}\delta^e_c$

remain notationally consistent when the labels are omitted from the diagrams. The "dimension" ν is depicted simply as a closed loop:

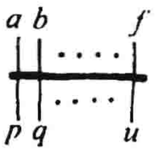


$= \delta^a_a = \nu.$

The utility of the diagrammatic notation is much enhanced when a special notation is employed for certain important elements of $\{\mathscr{X}\}$. Write



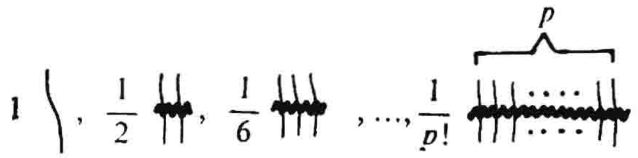
The general pattern should be clear. For each integer p we define two elements of $\{\mathscr{X}\}$. These have p arms and p legs and are sums (or differences) of $p!$ terms, one corresponding to each permutation of the legs (or, equivalently of the arms). Each term is an outer product of p "Kronecker deltas". The negative signs occur, in the second case, whenever the permutation is odd. (These are really Aitken's diagrams [1]. The permutation is odd if the number of intersection points between the lines is odd, counted correctly if multiple intersections are present.) In the case of ordinary tensors, the object



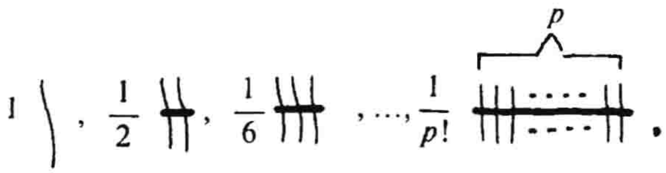
$$= \begin{vmatrix} \delta^a_p & \delta^b_p & \cdots & \delta^f_p \\ \delta^a_q & \delta^b_q & \cdots & \delta^f_q \\ \vdots & \vdots & & \vdots \\ \delta^a_u & \delta^b_u & \cdots & \delta^f_u \end{vmatrix} = \delta^{ab\ldots f}_{pq\ldots u}$$

will be recognized as the "generalized Kronecker delta".

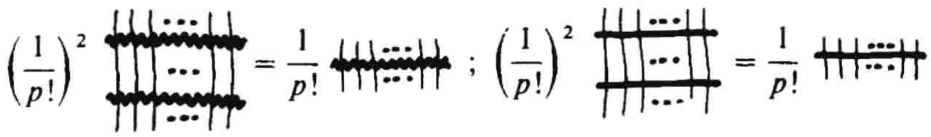
If the ring $\mathscr{T}$ of scalars contains (a subring isomorphic to) the rationals, then we can form the symmetrizers


1
$\frac{1}{2}$
$\frac{1}{6}$
, ...,
$\frac{1}{p!}$
p

and skew-symmetrizers

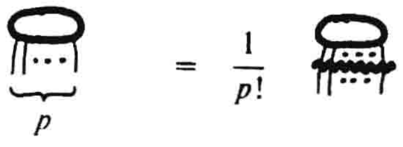

1
$\frac{1}{2}$
$\frac{1}{6}$
, ...,
$\frac{1}{p!}$
p

which have the idempotence:

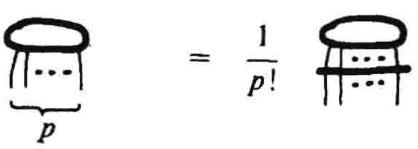

$\left(\frac{1}{p!}\right)^2$
$= \frac{1}{p!}$
;
$\left(\frac{1}{p!}\right)^2$
$= \frac{1}{p!}$

(As an alternative notation, we could absorb the $1/p!$ factors into the definitions of the diagrams. On the whole I prefer not to. This is partly because one would like to be able to handle the case when $\mathscr{T}$ has finite characteristic. But, in addition, it turns out that the factors $1/p!$ are rather a nuisance when expansions of complicated diagrams are carried out.) We can form symmetric and skew parts of tensors as follows:

symmetric part of

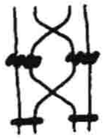


skew part of



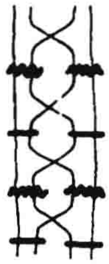
We can also build up Young tableau operators (cf. [7]) from these. For example, the expression

$$\frac{1}{12}$$

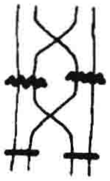


is an irreducible Young idempotent:

$$\left(\frac{1}{12}\right)^2$$

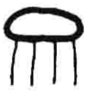


$$= \frac{1}{12}$$

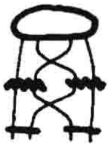


corresponding to the partition (2^2). When applied to a four-legged object:

$$\text{“}\begin{array}{|c|c|}\hline 1 & 3 \\ \hline 2 & 4 \\ \hline\end{array}\ \text{part” of}$$



$$= \frac{1}{12}$$



This particular operator produces an object with the symmetries of a Riemann–Christoffel tensor.

An ATS will be called *weakly regular* iff for all $\chi^{\mathscr{A}} \in \{\mathscr{T}\}$, there exists $\theta_{\mathscr{A}} \in \mathscr{T}_{\mathscr{A}}$ such that

$$(\chi^{\mathscr{A}}\theta_{\mathscr{A}})\chi^{\mathscr{B}} = \chi^{\mathscr{B}}$$

An ATS is *regular* iff for all $\chi^{\mathscr{A}\Gamma} \in \mathscr{T}^{::::}_{::::}$ (with every different possible clumping of its indices into two groups) there exists $\theta_{\Gamma\mathscr{A}}$ such that

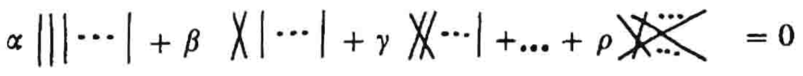
$$\chi^{\mathscr{A}\Gamma}\theta_{\Gamma\mathscr{C}}\chi^{\mathscr{C}\Delta} = \chi^{\mathscr{A}\Delta}.$$

(This terminology is suggested by the concept of a regular ring [8].) If we assume that the scalars $\mathscr{T}$ constitute a field, then clearly $\{\mathscr{T}\}$ is weakly regular iff for every non-zero $\chi^{\mathscr{A}} \in \{\mathscr{T}\}$, there exists $\theta_{\mathscr{A}} \in \mathscr{T}_{\mathscr{A}}$ such that

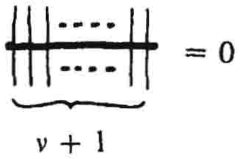
$$\chi^{\mathscr{A}}\theta_{\mathscr{A}} \neq 0.$$

Applications of negative dimensional tensors. *Combinatorial mathematics and its applications* (Proc. Conf., Oxford, 1969). Ed. D. J. A. Welsh, pp. 221–44

Furthermore, it can be shown [4] that if there is $\delta^b_a \in \mathscr{T}^b_a$ satisfying some identical relation



which is non-trivial (i.e. with some non-vanishing coefficient) and if $\{\mathscr{T}\}$ is weakly regular, with $\mathscr{T}$ a field, then $\{\mathscr{T}\}$ is regular. In fact, for any ATS whose elements are representable in the ordinary way in terms of (finite) sets of components and which contains δ^b_a, there will be an identical relation of this kind. For, the dimension $\nu = \delta^a_a$ must now be a positive integer, and we have



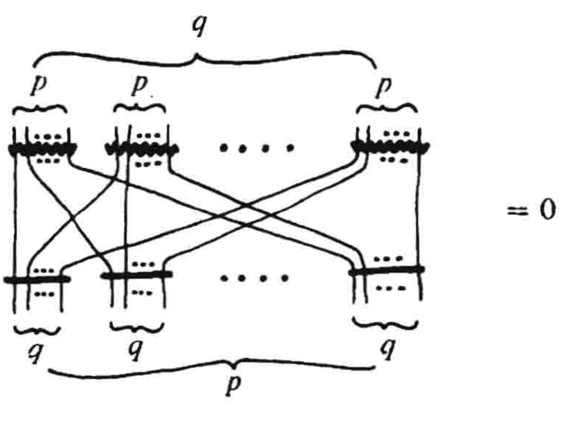
as a well-known simple property of the generalized Kronecker delta.

Simply for comparison with the weak regularity condition above, let us consider another (stronger), condition on an ATS (for which $\mathscr{T}$ is a field). We call an ATS *complete* if: for every non-zero $\chi^{\mathscr{A}\Gamma} \in \{\mathscr{T}\}$ (with each possible clumping into two clumps), there exists $\theta_{\mathscr{A}} \in \mathscr{T}_{\mathscr{A}}$ such that

$$\chi^{\mathscr{A}\Gamma}\theta_{\mathscr{A}} \neq 0.$$

Every such *complete* ATS will be isomorphic to the system of *all n*-dimensional ("ordinary") tensors over the field $\mathscr{T}$.

The possible $\{\mathscr{I}\}$-systems can be classified completely—$\mathscr{T} = \mathscr{I}$ being, for simplicity, a field without characteristic. The result (cf. [4]) is that either $\{\mathscr{I}\}$ is *free* (i.e. no identical relations of the type considered), or the identical relations are such that they can all be obtained from a single relation of the type



corresponding to a *rectangular* partition (p^q). We then have $\nu = q - p$. Call such a δ_a^b of type $[q-1, p-1]$ (p and q must be positive integers). The δ_a^b's which generate regular $\{\mathscr{I}\}$ systems are then just those of type $[n, 0]$ or $[0, n]$. A δ_a^b of general type $[m, n]$ is constructible as a "direct sum" of one of type $[m, 0]$ and one of type $[0, n]$. The "direct product" of δ_a^b of type $[m, n]$ with δ_α^β of type $[p, q]$ is δ_A^B of type $[mp + nq, mq + np]$ $\big(A = (a, \alpha), B = (b, \beta)\big)$.

A *Cartesian* ATS is one containing a special element $g_{ab} \in \mathscr{T}_{ab}$ which is symmetric:

$$g_{ab} = g_{ba}$$

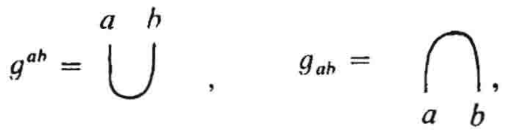
and invertible in the sense that there exists another element $g^{ab} \in \mathscr{T}^{ab}$ with

$$g_{ab}\, g^{bc} = \delta_a^c$$

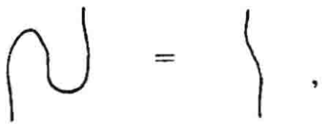
(Since $g_{cd}\, g^{ca}\, g^{db} = g^{ab}$, it follows that g^{ab} is also symmetric: $g^{ab} = g^{ba}$.) The elements g_{ab}, g^{ab} can be used, in the usual way, to translate an upper index into a lower one, or a lower index into an upper one:

$$\chi_{ad}{}^{c} = \chi_{a}{}^{bc}\, g_{bd}, \; \chi_{a}{}^{bc} = \chi_{ad}{}^{c}\, g^{db}.$$

It now ceases to be important to maintain a distinction between upper and lower indices. For, writing



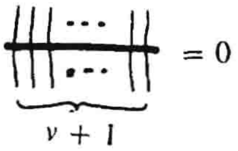
which is consistent with



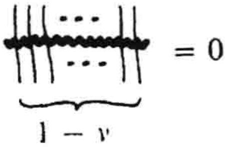
we can turn arms into legs and vice-versa. Since it now makes no difference whether an arm points up or down, we may as well also allow arms to point right or left or, indeed, in any direction in the plane. This will give our diagrams an additional flexibility. (In fact we can attain this flexibility also in the case of a non-Cartesian system if we allow our index lines to be *directed*—as may be indicated, say, by an arrow on the line, rather than simply by the vertical ordering.)

The complete description of Cartesian $\{\overset{c}{\mathscr{I}}\}$-systems (i.e. systems generated by g_{ab}, g^{ab} and scalars) is a little more complicated than that of the non-Cartesian $\{\overset{c}{\mathscr{I}}\}$-systems. Assume $\mathscr{I}$ is a field without characteristic, as before.

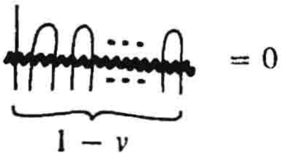
Then there is a unique regular $\{\overset{c}{\mathscr{J}}\}$-system for each integral value of ν. When $\nu = 2, 3, 4, \ldots$ (or in the trivial cases $\nu = 0, 1$) we have simply a system isomorphic with the "ordinary" Cartesian system of (numerically) invariant tensors over a ν-dimensional vector space, with component representation in the normal way. We have the basic identical relations:



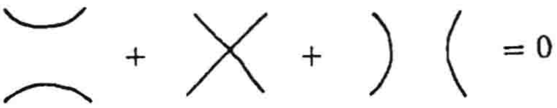
with δ_a^b of type $[\nu, 0]$. When $\nu = -2, -4, -6, \ldots$ we have δ_a^b of type $[0, -\nu]$ with



as an identical relation. However, this can be contracted down to give a relation of lower valence:

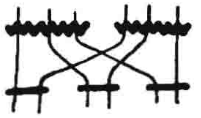


The most interesting system of all is, perhaps, the regular Cartesian system with $\nu = -2$. The elements of this system I call *binors*. The above basic identical relation can be written out as



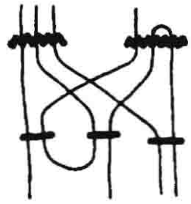
We readily see that every contraction of this expression is identically zero if $v = -2$. This expression must indeed vanish for a (weakly) regular $\{\overset{c}{\mathscr{J}}\}$-system with $\nu = -2$.

When $\nu = -1, -3, -5, \ldots$, the systems are somewhat more complicated. The type is now $[1, 1-\nu]$ instead of $[0, -\nu]$. Let us illustrate this with the case $\nu = -1$. We have the identical relation



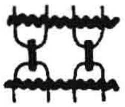
$= 0$

but this contracts down to give (equivalent) relations of smaller valence, e.g.

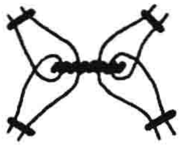


$= 0.$

More symmetrical forms, *each* of which is equivalent to either of the above two relations, are

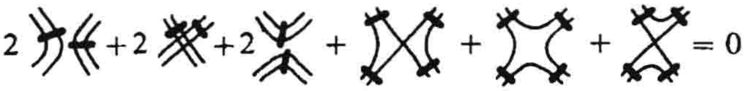


$= 0,$



$= 0.$

This second expression, partly expanded out, is:



We readily verify that every contraction of this vanishes if $\nu = -1$. The cases $\nu = -3, -5, \ldots$ are essentially similar.

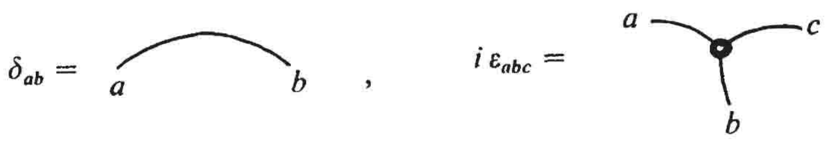
So much for general theory. Now let us examine some applications.† First consider a simple (regular) Cartesian system generated by the complex numbers as scalars and an ordinary three-dimensional Cartesian Kronecker delta δ_{ab} $(= g_{ab})$ and the skew-symmetrical (Levi-Civita) ε_{abc}. We can represent δ_{ab} and ε_{abc} in the normal way in terms of components, with

$$\delta_{11} = \delta_{22} = \delta_{33} = 1,$$
$$\varepsilon_{123} = \varepsilon_{312} = \varepsilon_{231} = 1,$$
$$\varepsilon_{132} = \varepsilon_{321} = \varepsilon_{213} = -1,$$

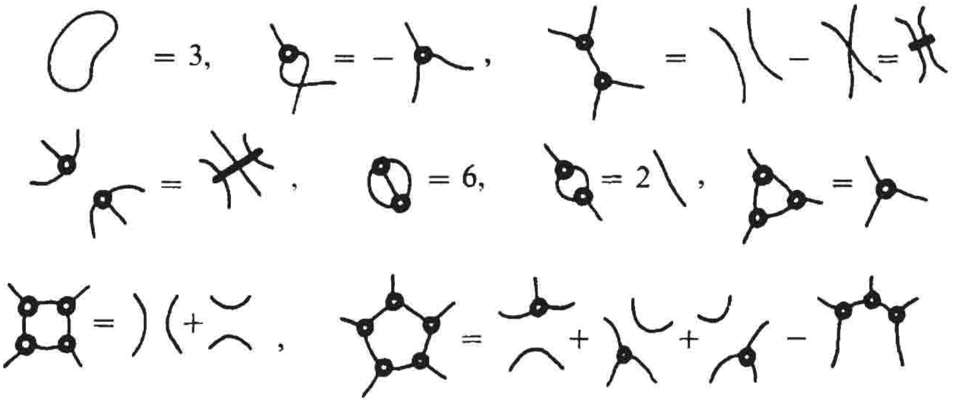
---

† The main application described here was, for the most part, obtained during the period 1953–55, but has not been published hitherto.

the remaining components being zero. Contraction is represented by the Einstein summation convention in the usual way. Depict:

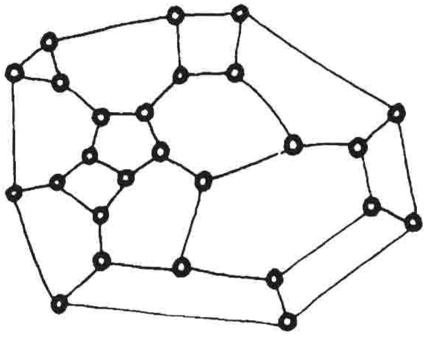


where the factor $i\,(=\sqrt{-1})$ is included for simplicity in the signs of reduction formulae. We have, as well-known formulae of tensor calculus, or by repeated applications, thereof:



etc.†

Now consider a planar graph of degree three. We can associate with it a certain contracted product of ε_{abc}'s, where one $i\varepsilon_{abc}$ is drawn at each vertex, a contraction occurring for each edge of the graph. The result is just some complex number (actually an integer). For example:

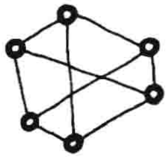

$= K.$

The number K is, in fact, precisely the number of ways of colouring the edges of the graph with three colours so that three distinct colours occur at each vertex. This, by a well-known result [2], is just one-quarter of the number of ways of colouring the faces of the graph with four colours so that no two similarly coloured faces have an edge in common.

---

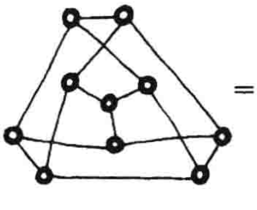
† I am grateful to T. G. Murphy for pointing the last of these formulae out to me.

To see the connection between K and the edge-colouring number, we think of the representation in terms of components. The tensor expression is then simply a sum of terms, each of which arises when one of the numbers 1, 2, 3 is assigned to each edge (corresponding to a summed index). We get a factor i for each vertex where the numbers 1, 2, 3 appear cyclicly, and $-i$ if 1, 2, 3 appear anti-cyclicly. The term is zero if a repetition of 1, 2, 3 appears at some vertex. Thus each non-zero term arises from an allowable colouring of the edges with three colours. Finally we have to check that each non-zero term contributes precisely the value $+1$. It is at this point that the planarity of the graph enters. There are various ways of seeing that the sign comes out correctly here, but I have not been able to think of an argument which can be presented in a nutshell, so I shall just omit it here. It is, however, worthwhile to illustrate the necessity of the planarity condition by an example. We have



$= 0$

even though the graph's edges *can* be coloured with three colours (in twelve ways). The vanishing of the tensor expression follows from the skew-symmetry of ε_{abc}. For, a pair of vertices can be interchanged while leaving the configuration invariant, but with three signs changed in the process. Of course any non-planar graph of degree three whose edges are *not* colourable with three colours *must* have a vanishing tensor expression, e.g. the Petersen graph gives:



$= 0$

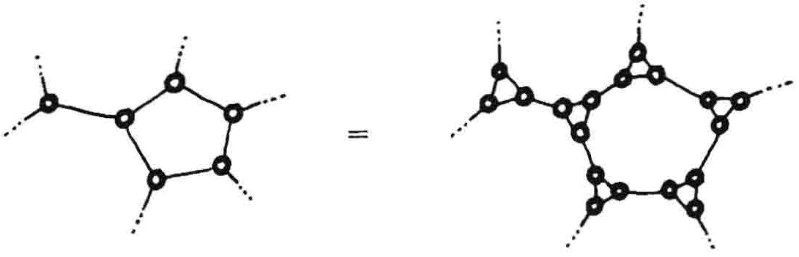
(the vanishing is also a consequence of skew-symmetry again in this case).

There is an alternative way of establishing the equality of the tensor expression K and the number of colourings in the case of a planar graph. This is simply to observe that the last four of our reduction formulae for ε_{abc} expressions can be used to eliminate the faces one by one. We merely have to verify that these four reduction formulae are equally valid if they refer to numbers of colourings or to ε_{abc} expressions. Then we finally end up with a

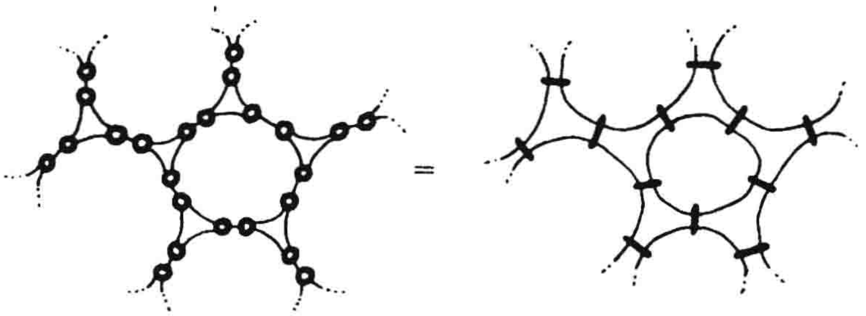
Applications of negative dimensional tensors. *Combinatorial mathematics and its applications* (Proc. Conf., Oxford, 1969). Ed. D. J. A. Welsh, pp. 221–44

single closed loop. This "edge" can be coloured in three ways and the value is 3, thus completing the induction.

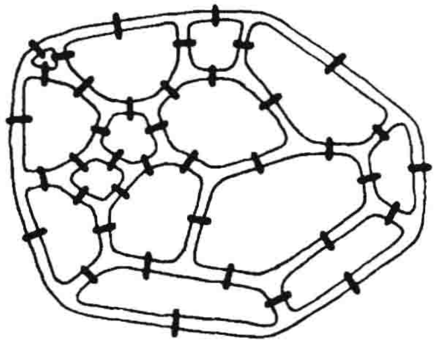
Now let us try something further. We can replace each vertex by a little triangle, without changing the value (because of one of our reduction formulae). Thus



The two new vertices at the ends of each original edge can now be slid together and another of our reduction formulae employed to give:

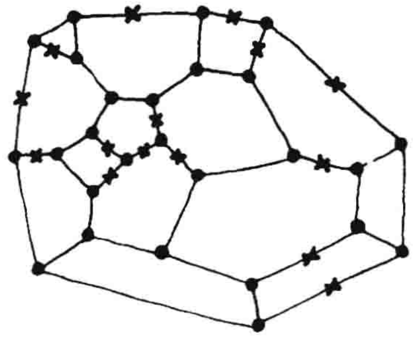


Thus, we have a new expression for K:

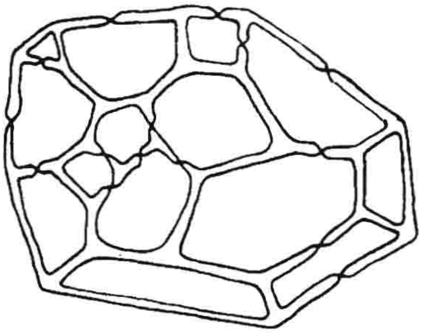


$= K.$

We can reinterpret this combinatorially. Expanding out each of the "skew-symmetrizers" we get a sum of 2^E terms, where E is the number of edges in the graph. We can describe each term as follows. Select a subset of edges of the graph and call these the crossed edges. Then the term is positive or negative according as the number of crossed edges is even or odd. We build up a number of closed loops by following along just inside the boundary of each face and then crossing over to the other side of the edge at its mid-point whenever the edge is a crossed edge. Let the total number of closed loops thus formed be C. Then the absolute value of the term is 3^C, (since a closed loop takes the value $v = 3$). For example, in the graph just considered, a typical term is given by an assignment of crossed edges as follows:



The corresponding term is

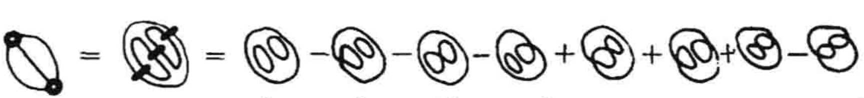
$-$  $= -3^5.$

since there are five closed loops, the minus sign occurring because there are 15 crossed edges. Thus we have

$$K = \sum (-1)^X 3^C$$

where X is the number of crossed edges and C the resulting number of closed loops.

A simple worked example is the following:



$$= 3^3 \quad - 3^2 \quad - 3^2 \quad - 3^2 \quad + 3 \quad + 3 \quad + 3 \quad - 3 \quad = 6.$$

This is obviously the correct answer, but it also illustrates the fact that we must expect a considerable amount of cancellation in the sum. The formula is clearly not very practical for explicit calculation, but also it is unlikely to be of

Applications of negative dimensional tensors. *Combinatorial mathematics and its applications* (Proc. Conf., Oxford, 1969). Ed. D. J. A. Welsh, pp. 221–44

great direct theoretical use in attempts to establish $K > 0$, owing to this cancellation.

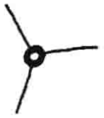
At this point we turn our attention to a "negative dimensional" ATS. It is a remarkable fact that the above formula for K has a curious analog in which 3 is replaced by -2:

$$K = (-\tfrac{1}{4})^{\frac{1}{2}V} \sum (-1)^X (-2)^C$$

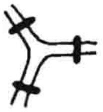
where $V = \frac{2}{3}E$ is the number of vertices of the graph. Note that the value of the summation must now be much larger, even though the individual terms have smaller absolute value than before. Thus the amount of cancellation has been considerably reduced. The above worked example gives:

$$\begin{aligned} K &= -\tfrac{1}{4}\{(-2)^3 - (-2)^2 - (-2)^2 - (-2)^2 \\ &\qquad\qquad + (-2) + (-2) + (-2) - (-2)\} \\ &= \tfrac{1}{4}(8 + 4 + 4 + 4 + 2 + 2 + 2 - 2) = 6. \end{aligned}$$

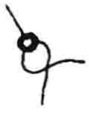
The basis of the above formula stems from consideration of the system of binors (i.e. the regular system $\{\overset{c}{\mathscr{I}}\}$ with $\nu = -2$). We consider a correspondence



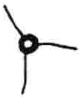
$$\leftrightarrow \frac{i}{4\sqrt{2}}$$

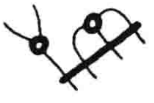


where the arms on the left-hand figure are ordinary Cartesian 3-tensor lines (as before, i.e. $\nu = 3$), but where the arms on the right-hand figure are binor lines ($\nu = -2$). The point about this correspondence is that the algebraic properties of the $i\varepsilon_{abc}$ object on the left and of the binor object on the right—indices suitably clumped in pairs as indicated—are formally identical. In other words, the correspondence establishes an embedding of the $(\delta_{ab}, \varepsilon_{abc}; \nu = 3)$ system, considered earlier, in the binor system $(\delta_{\alpha\beta}; \nu = -2)$ for which indices are clumped in pairs (e.g. $a = (\alpha, \alpha_0)$, $b = (\beta, \beta_0)$, ...) I shall not go into the matter in full detail here. Essentially, what has to be verified is that the relations

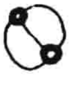


$= -$



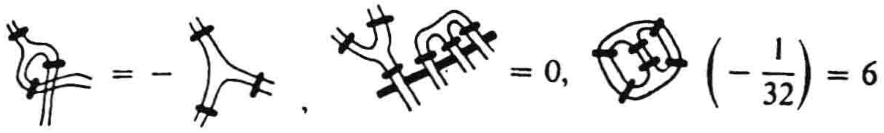


$= 0.$

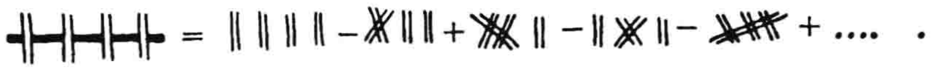


$= 6$

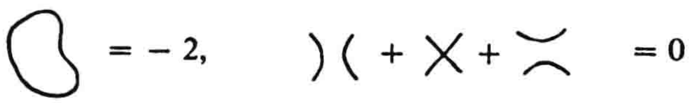
are also satisfied by the proposed binor object. That is,



where

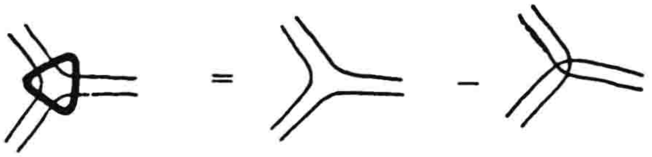


The verification of these relations requires the properties

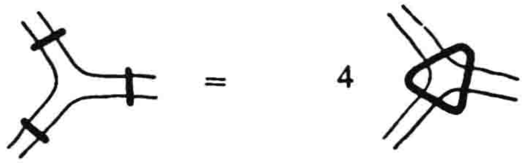


which characterize the binor system. It is easiest not to carry out the fu verification of the middle relation directly, but to use some general result: These will not be entered into here. The third relation is really just the exampl considered earlier.

By using binors, we can also establish some other formulae for *K*. Set

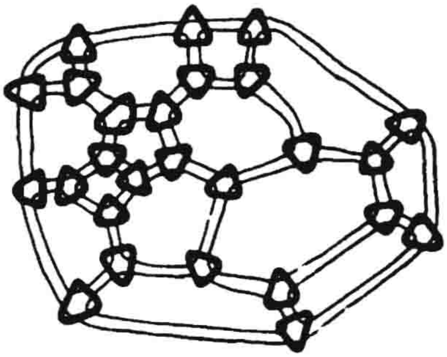


Then we have the binor relation



This may be verified directly using the basic binor identical relation. Thi

$$K = \left(-\frac{1}{2}\right)^{\frac{1}{2}V}$$



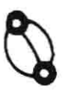
If we assign an orientation "*counter-clockwise*" to each face, we may describe this formula in terms of *directed* circuits (i.e. this falls within the framework of the non-Cartesian system $\{\mathscr{I}\}$ with $v = -2$). (We could not do this previously because the orientations would not have been consistent with crossed edges.) Then our formula can be described as follows:

$$K = (-\tfrac{1}{2})^{\frac{1}{2}V} \sum (-1)^Y (-2)^{C'}$$

where there are 2^V terms in the sum, one term corresponding to each assignment of a "+" or of a "−" to every vertex. Here Y is the number of "−" vertices; C' is the number of directed circuits required, where we cover every edge of the graph exactly twice and where we turn to the *left* at every "+" vertex and to the *right* at every "−" vertex, as we follow the circuits. In fact, it can be seen fairly easily that the number of circuits has the same parity in each term. Hence

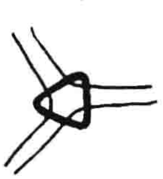
$$K = (\tfrac{1}{2})^{\frac{1}{2}V} \sum (-1)^Y 2^{C'}$$

(So apart from an overall sign $(-1)^{\frac{1}{2}V}$, we may interpret our last diagram for K as applying within the system $\{\mathscr{I}\}$ or $\{\overset{c}{\mathscr{I}}\}$ with $v = 2$.) Our example now gives:

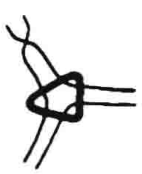


$$= K = \tfrac{1}{2}\{8 - 2 - 2 + 8\} = 6.$$

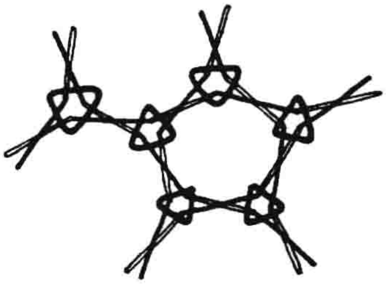
Let us consider one final expression for K. We observe that our binor vertex possesses the skew-symmetry



$$= \; -$$



(since we saw that it was equal to an expression in which this skew-symmetry was manifest). Thus, at most the overall sign is altered if we replace our last diagram for K by one in which all edges become crossed edges. If we depict the diagram in the following way,



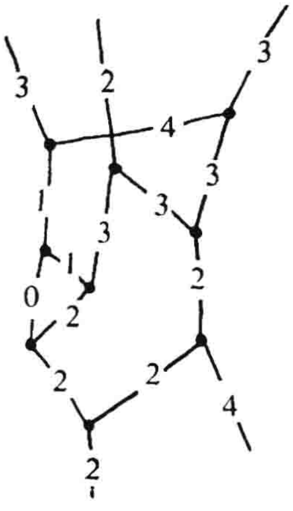
then an alternative formula for K presents itself:

$$K = (\tfrac{1}{2})^{\frac{1}{2}V} \sum (-1)^{Y} (-2)^{C''}.$$

Here, as before, there are 2^V terms in the sum; each term corresponds to an assignment of a "+" or a "−" to every vertex; Y is the number of "−" vertices. Now, C'' is the number of (undirected) closed circuits, consisting of alternately black and white line segments, which just cover every edge twice, once with a black segment and once with a white segment, where at each "+" vertex we turn ***right*** from white to black and ***left*** from black to white and *vice-versa* at each "−" vertex.

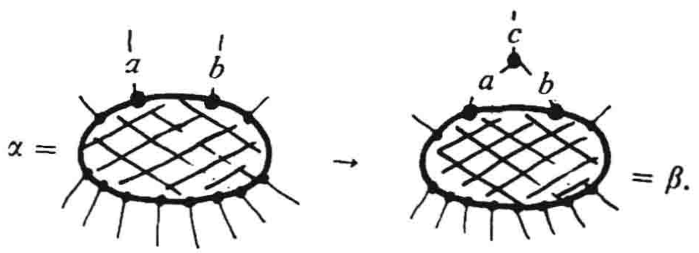
There is another application of binor algebra which is worth mentioning here. This is to the theory of quantum mechanical angular momentum. I shall not go into much detail here. A more extended account has been given elsewhere [6].

Consider an undirected graph of degree 3, but which may possess also some free ends. Each edge of the graph (including each free end) is assigned a non-negative integer. I suppose that the sum of the three integers at each vertex is *even* and that no integer at a vertex exceeds the sum of the other two. I call such a graph a *spin-network*. An example of a spin-network is the following:



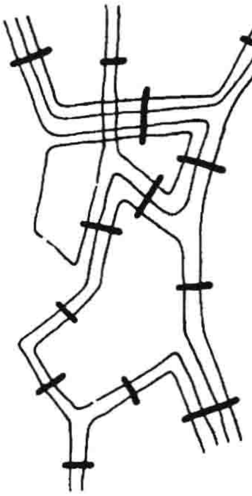
Each spin-network will have a physical interpretation—or, rather, a number of different physical interpretations, depending on a choice of "time-ordering" for the graph. Let us think of the network as a kind of space–time diagram of a portion of the universe. For the spin-network depicted above, we shall think of time as progressing from the bottom of the diagram to the top. Each line segment is to be pictured as describing a part of the universe—called a *unit*—which is effectively stationary and isolated from the remainder of the universe (e.g. a "free particle"). I suppose that a unit also has a well-defined total angular momentum. According to quantum mechanics, the value of this total angular momentum must be a (non-negative) *integral* multiple of $\frac{1}{2}\hbar$. If the integer depicted on the line segment is n, then the total angular momentum of the unit is to be $\frac{1}{2}n\hbar$ and the unit is called an *n-unit*.

We may interpret the above diagram as follows: A 2-unit splits into two 2-units one of which combines with a 4-unit to make another 2-unit; the other 2-unit splits into a 0-unit and yet another 2-unit, which subsequently splits into a 1-unit and a 3-unit, the 1-unit combining with the 0-unit to make another 1-unit; ...... We are interested in the following type of question. Suppose we are given a spin-network in which an a-unit and a b-unit emerge. Suppose that the a-unit and the b-unit combine to form a c-unit:



We ask: what is the probability value for each possible value of the integer c, if the remainder of the spin-network is all that we are given? Now, quantum mechanics gives a well-defined answer to this question (assuming the effect of relative motion between the different units may be neglected). There is a standard procedure for obtaining this probability, but calculations tend to get very involved.

An alternative (but equivalent) procedure is to use binors. With each spin-network we associate a binor, where an n-unit is represented by n lines with a "skew-symmetrizer" across. The lines are then joined up at each vertex in an essentially unique way. For example, the above spin-network is represented as

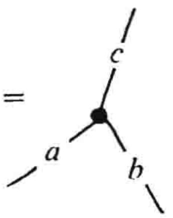


Define a norm $||\kappa||$, for a spin-network κ, to be the modulus of the scalar obtained by joining together two identical copies of the binor for κ, connecting contracting over corresponding free arms. Then, my assertion is that when a-unit and the b-unit combine in the above process, the probability that the resulting unit is a c-unit is:

$$\text{probability of } c = \frac{\|\beta\|\,(c+1)(a!)^2(b!)^2}{\|\alpha\|\;\|\gamma\|}$$

where

$\gamma =$



It may be finally remarked that although the *direct* calculation of spin-network norms, by the above prescription, can be very impractical, there are various shortcuts possible. One such method† involves counting the total number of ways of colouring a certain series of graphs, associated with the spin-network, with n-colours and then putting $n = -2$ in the resulting polynomial. In this way it is possible to obtain some standard angular momentum formulae (e.g. Racah's expression for the $6-j$ symbol) very rapidly. All this will be described fully elsewhere.

† When applied to the particular case of a planar graph (no free ends and all spin values having $n = 2$), this method yields yet *another* expression for K. Consider each different way of covering the edges of the graph with a number of circuits, where each edge is covered

References

1. A. C. Aitken (1958). "Determinants and Matrices". Oliver & Boyd, Edinburgh.
2. P. J. Heawood (1897). *Quart. J. of Math.* (Oxford). **29,** 270–285.
3. T. G. Murphy, "Tensor" (duplicated manuscript, Trinity College Dublin; to be published in book form later).
4. R. Penrose (1956). "Tensor Methods in Algebraic Geometry" (Ph.D. Thesis, Cambridge University).
5. R. Penrose (1968). Structure of space-time. *In* "Battelle Recontres" (1967 Lectures in Mathematics and Physics) (Eds. C. M. De Witt and J. A. Wheeler), W. A. Benjamin, Inc., New York.
6. R. Penrose (1969). Angular Momentum: an approach to Combinatorial Space-Time. *In* "Quantum Theory and Beyond" (Ed. E. A. Bastin), Cambridge University Press, Cambridge.
7. D. E. Rutherford (1948). "Substitutional Analysis", Hafner.
8. J. von Neumann (1936). On regular rings. *Proc. Nat. Acad. Sci.* **22,** 707–713.

---

exactly twice and where no individual circuit covers any edge more than once. For each such covering, colour the circuits with ν colours so that no two similarly coloured circuits cover a common edge. Count the number of distinct colourings for each such covering and sum over all possible such coverings. The result will be a polynomial in ν. Substituting the value $\nu = -2$ into this polynomial we get $(-4)^{V}{\tfrac{1}{4}}K$. (If we substitute $\nu = 3$ into the polynomial, the result is in fact also just K, but this is less subtle, being quite easy to see directly.)

The above polynomial in ν can also be obtained in another way. We assign a string of ν symbols, consisting solely of "zeros" and "ones", to every *face* of the graph (counting the external "sea" as a face). The strings of symbols assigned to two adjacent faces are to differ in exactly *two* places. The number of ways of making such an assignment is then just 2^{ν} times the required polynomial.

46

ANGULAR MOMENTUM: AN APPROACH TO COMBINATORIAL SPACE-TIME

Quantum Theory and Beyond. Ed. Ted Bastin, Cambridge University Press, 1971

This was really the first reasonably comprehensive account of spin-network theory (see Chapter 31) that I had actually completed, but it also appears to provide the first announcement, on p. [28], of the physical interpretation of a twistor Z^α as describing the momentum and angular momentum of a massless particle, its *helicity* being $\frac{1}{2}Z^\alpha\bar{Z}_\alpha$. Moreover, it seems to provide the first appearance of the twistor quantum commutator $Z^\alpha\bar{Z}_\beta - \bar{Z}_\beta Z^\alpha = \delta^\alpha_\beta\hbar$.

ANGULAR MOMENTUM: AN APPROACH TO COMBINATORIAL SPACE-TIME

ROGER PENROSE

I want to describe an idea which is related to other things that were suggested in the colloquium, though my approach will be quite different. The basic theme of these suggestions have been to try to get rid of the continuum and build up physical theory from discreteness.

The most obvious place in which the continuum comes into physics is the structure of space-time. But, apparently independently of this, there is also another place in which the continuum is built into present physical theory. This is in quantum theory, where there is the superposition law: if you have two states, you're supposed to be able to form any linear combination of these two states. These are complex linear combinations, so again you have a continuum coming in—namely the two-dimensional complex continuum—in a fundamental way.

My basic idea is to try and build up both space-time and quantum mechanics simultaneously—from *combinatorial* principles—but not (at least in the first instance) to try and change physical theory. In the first place it is a *reformulation*, though ultimately, perhaps, there will be some changes. Different things will suggest themselves in a reformulated theory, than in the original formulation. One scarcely wants to take *every* concept in existing theory and try to make it combinatorial: there are too many things which look continuous in existing theory. And to try to eliminate the continuum by *approximating* it by some discrete structure would be to change the theory. The idea, instead, is to concentrate only on things which, in fact, *are* discrete in existing theory and try and use them as *primary concepts*—then to build up other things using these discrete primary concepts as the basic building blocks. Continuous concepts could emerge in a limit, when we take more and more complicated systems.

The most obvious physical concept that one has to start with, where quantum mechanics says something is discrete, and which is connected with the structure of space-time in a very intimate way, is in *angular momentum*. The idea here, then, is to start with the concept of angular momentum—here one has a *discrete* spectrum—and use the rules for combining angular

---

This paper originally appeared in *Quantum Theory and Beyond*, edited by Ted Bastin, Cambridge University Press 1971, pp. 151–180.

momenta together and see if in some sense one can construct the concept of *space* from this.

One of the basic ideas here springs from something which always used to worry me. Suppose you have an electron or some other spin $\frac{1}{2}\hbar$ particle. You ask it about its spin: is it spinning up or down? But how does it know which way is 'up' and which way is 'down'? And you can equally well ask the question whether it spins right or left. But whatever question you ask it about directions, the electron has only just *two* directions to choose from. Whether the alternatives are 'up' and 'down', or 'right' and 'left' depends on how things are connected with the macroscopic world.

Also you could consider a particle which has *zero* angular momentum. Quantum mechanics tells us that such a particle has to be spherically symmetrical. Therefore there isn't really any choice of direction that the particle can make (in its own rest-frame). In effect there *is* only one 'direction'. So that a thing with zero angular momentum has just one 'direction' to choose from and with spin one-half it would have *two* 'directions' to choose from. Similarly, with spin one, there would always be just three 'directions' to choose from, etc. Generally , there would be $2s + 1$ 'directions' available to a spin s object.

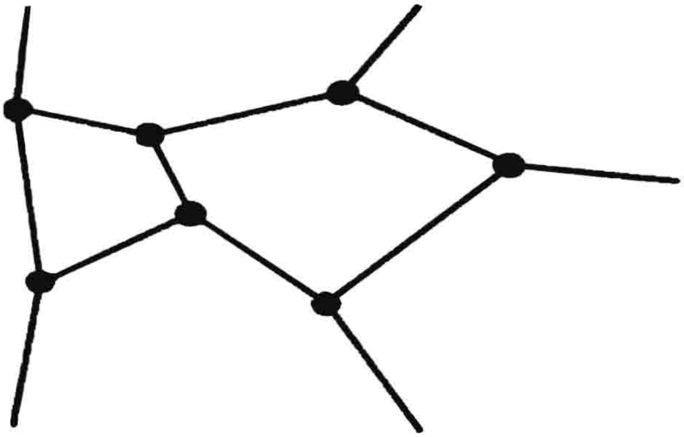
Of course I don't mean to imply that these are just directions in space in the ordinary sense. I just mean that these are the choices available to the object as regards its state of spin. That is, however we may choose to interpret the different possibilities when viewed on a macroscopic scale, the object itself is 'aware' only that these are the different possibilities that are open to it. Thus, if the object is in an s-state, there is but one possibility open to it. If it is in a p-state there are three possibilities, etc., etc. I don't mean that these possibilities are things that from a macroscopic point of view we would necessarily think of as directions in all cases. The s-state is an example of a case where we would not!

So we oughtn't at the outset to have the concept of macroscopic space-direction built into the theory. Instead, we ought to work with just these discrete alternatives open to particles or to simple systems. Since we don't want to think of these alternatives as referring to pre-existing directions of a background space—that would be to beg the question—we must deal only with *total* angular momentum (j-value) rather than spin in a direction (m-value).

Thus, the primary concept here has to be the concept of *total angular momentum* not the concept of angular momentum in, say, the z-direction, because: which is the z-direction?

Imagine, then, a universe built up of things like that shown in fig. 1. These lines may be thought of as the world-lines of particles. We can view *time* as going in one direction, say, from the bottom of the diagram to the top. But it turns out, really, that it's irrelevant which way time is going. So I don't want to worry too much about this.

I'm going to put a number on each line. This number, the *spin-number* will have to be an integer. It will represent twice the angular momentum, in units of $\hbar$. All the information I'm allowed to know about this picture will be just this diagram (fig. 2): the network of connections and spin numbers 3, 2, 3, ... like that. I should say



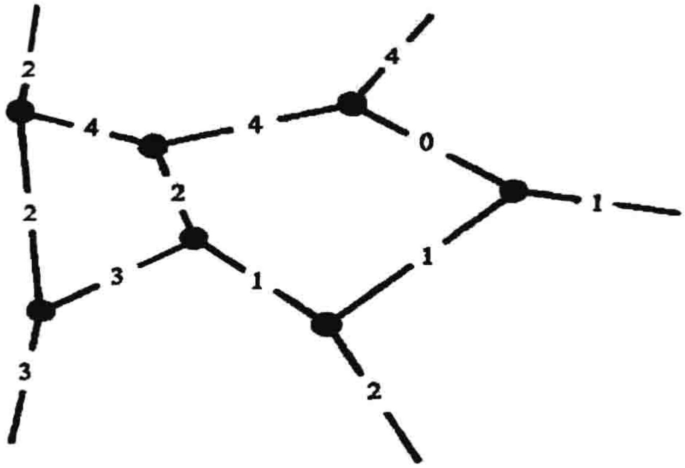


Fig. 2

that the picture I want to give here is just a model. Although it does describe a type of *idealized* situation exactly according to quantum theory, I certainly don't want to suggest that the universe 'is' this picture or anything like that. But it is not unlikely that *some* essential features of the model that I am describing could still have relevance in a more complete theory applicable to more realistic situations.

I have referred to these line segments as representing, in some way, the world-lines of particles. But I don't want to imply that these lines stand just for *elementary* particles (say). Each line could represent some compound system which separates itself from other such systems for long enough that (in some sense) it can be regarded as isolated and stationary, with a well-defined total angular momentum $n \times \frac{1}{2}\hbar$. Let us call such a system or particle an n-unit. (We allow $n = 0, 1, 2, \ldots$) For the precise model I am describing, we must also imagine that the particles or systems are not moving relative to one another. They just transfer angular momentum around, regrouping themselves into different subsystems, perhaps annihilating one another, perhaps producing new units. In the diagram (fig. 2), the 3-unit at the bottom on the left splits into a 2-unit and another 3-unit. This second 3-unit combines with a 1-unit (produced in the break up of a 2-unit into two 1-units) to make a new 2-unit, etc., etc. It is only the *topological* relationship between the different segments, together with the spin-number values, which is to have significance. The time-ordering of events will actually play no role here (except conceptually). We could, for example, read the diagram as though time increased from the left to the right, rather than from the bottom to the top, say.

Angular momentum conservation will be involved when I finally give the rules for these diagrams. These rules, though combinatorial, are actually derived from the standard quantum mechanics for angular momentum. Thus, in particular, the conservation of total angular momentum must be built into the rules.

Now, I want to indicate answers to two questions. First, what *are* these combinatorial rules and how are we to interpret them? Secondly, how does this enable us to build up a concept of space out of total angular momentum? In order not to get bogged down at this stage with too much detail, I shall defer, until later on, the complete definition of the combinatorial rules that will be used. All I shall say at this stage is that every diagram, such as fig. 2 (called a *spin-network*) will be assigned a non-negative integer which I call its *norm*. In some vague way, we are to envisage that the norm of a diagram gives us a measure of the frequency of occurrence of that particular spin-network in the history of the universe. This is not actually quite right—I shall be more precise later—but it will serve to orient our thinking. We shall be able to use these norms to calculate the probabilities of various spin values occurring in certain simple 'experiments'. These probabilities will turn out always to be rational numbers, arising from the fact that the norm is always an integer. Given any spin-network, its norm can be calculated from it in a purely combinatorial way. I shall give the rule later.

But first let me say something about the answer to the second question. How can I say anything about *directions* in space, when I only have the non-directional concept of total angular momentum? How do I get 'm-values' out of j-values, in other words?

Clearly we can't do quite this. In order to know what the 'm-value' of an n-unit is, we would require knowledge of which direction in space is the 'z-direction'. But the 'z-direction' has no physical meaning. Instead, we may ask for the 'orientation' of one of our n-units *in relation to some larger structure* belonging to the system under consideration. We need some larger structure which in fact does give us something that we may regard as a well-defined 'direction in space' and which could serve in place of the 'z-direction'. As we have seen, a structure of spin zero, being spherically symmetrical, is no good for this; spin $\frac{1}{2}\hbar$ is not much better; spin $\hbar$ only a little better; and so on. Clearly we need a system involving a fairly large total angular momentum number if we are to obtain a reasonably well-defined 'direction' against which to test the 'spin direction' of the smaller units. We may imagine that for a large total angular momentum number N, we have the potentiality, at least, to define a well-defined direction as the *spin axis* of the system. Thus, if we *define* a 'direction' in space as something associated with an N-unit with a large N value (I call this a *large unit*), then we can ask how to define *angles* between these 'directions'. And if we can decide on a good way of measuring angles, we can then ask the question whether the angles we get are consistent with an interpretation in terms of directions in a Euclidean three-dimensional space, or perhaps in some other kind of space.

How, then, are we to define an angle between two large units? Well, we do this by performing an 'experiment'. Suppose we detach a 1-unit (e.g. an electron, or any other spin $\frac{1}{2}\hbar$-particle) from a large N-unit in such a way as to leave it as an $(N-1)$-unit. We can then re-attach the 1-unit to some other large unit, say an M-unit. What does the M-unit do? Well (according to the rules we are allowed here) it can either become an $(M-1)$-unit or an $(M+1)$-unit. There will be a certain probability of one outcome and a certain probability of the other. Knowing these probability values, we shall have information as to the angle between the N-unit and the M-unit. Thus, if our two units are to be 'parallel', we would expect zero probability for the $M-1$ value and certainty for the $M+1$ value. If the two units are to be 'anti-parallel' we would expect exactly the reverse probabilities. If they are 'perpendicular', then we would expect equal probability values of $\frac{1}{2}$, for each of the two outcomes. Generally, for an angle θ between the directions of the two large units we would expect a

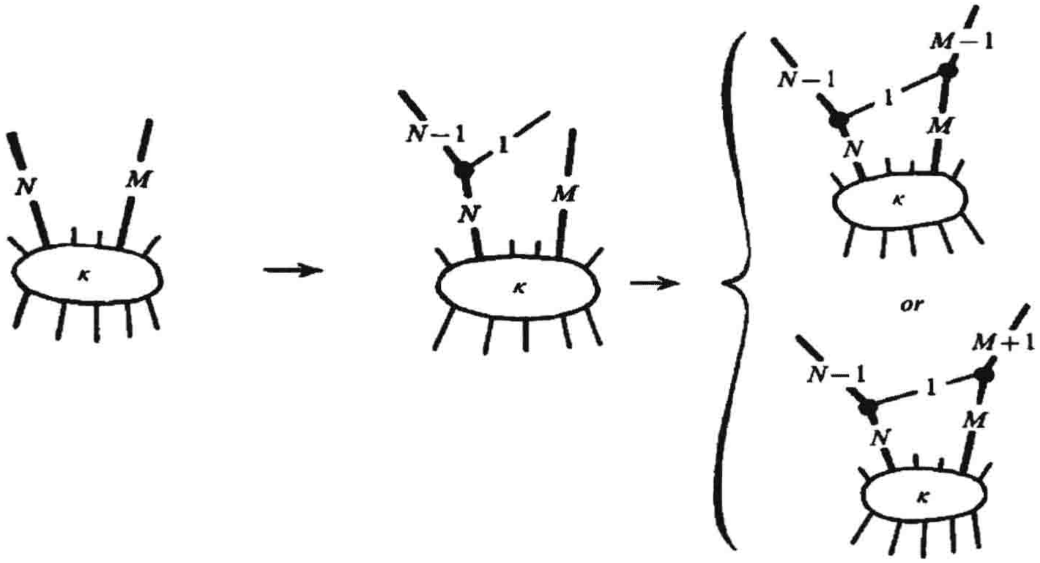


Fig. 3

probability $\frac{1}{2} - \frac{1}{2}\cos\theta$ for the M-unit to be reduced to an $(M-1)$-unit and a probability $\frac{1}{2} + \frac{1}{2}\cos\theta$ for it to be increased to an $(M+1)$-unit. Let me draw a diagram to represent this experiment (fig. 3). Here κ represents some known spin-network. By means of a precise (combinatorial) calculational procedure—which I shall describe shortly—we can calculate, from knowledge of the spin-network κ, the probability of each of the two possible final outcomes. Hence, we have a way of getting hold of the concept of Euclidean angle, starting from a purely combinatorial scheme.

As I remarked earlier, these probabilities will always turn out to be *rational* numbers. You might think, then, that I could only obtain angles with rational cosines in this way. But this would be a somewhat misleading way of viewing the situation. With a finite spin-network with finite spin-numbers, the angle can never be quite well-enough defined. I *can* work out numerical values for these 'cosines of angles' for a finite spin-network, but these 'angles' would normally not quite agree with the actual angle of Euclidean space until I go to the limit.

The view that I am expressing here is that rational probabilities are to be regarded as something which can be more primitive than ordinary real number probabilities. I don't need to call upon the full continuum of probability values in order to proceed with the theory. A rational probability $p = m/n$ might be thought of as arising because the universe has to make a choice between m alternative possibilities of one kind and n alternative possibilities of another—all of which are to be equally probable. Only in the limit, when numbers go to infinity do we expect to get the full continuum of probability values.

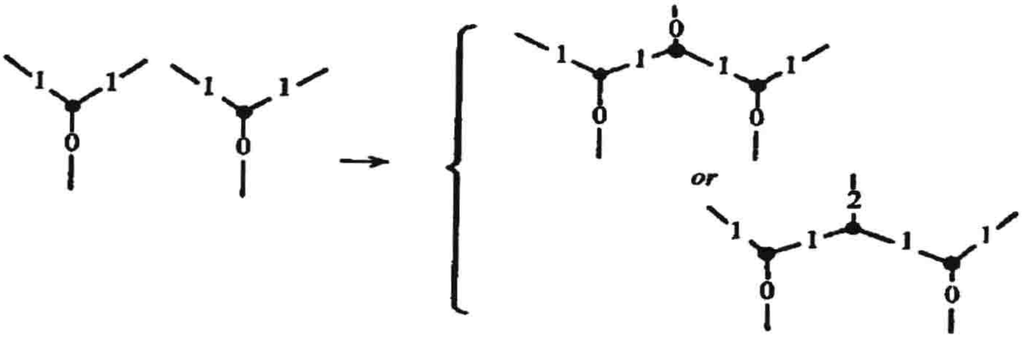


Fig. 4

As a matter of fact, it was this question of rational values for primitive probabilities arising in nature, which really started me off on this entire line of thought concerning spin-networks, etc. The idea was to find some situation in nature which one might reasonably regard as giving rise to a 'pure probability', I am not really sure whether it is fair to assume that 'pure probabilities' exist in nature, but by these I mean probabilities (necessarily quantum mechanical) whose values are determined by nature alone and not in principle influenced by our ignorance of initial conditions, etc. I suppose I might have thought of branching ratios in particle decays as a possible example. Instead, I was led to consider a situation of the following type.

Two spin zero particles each decay into pairs of spin $\frac{1}{2}\hbar$ particles. Two of the spin $\frac{1}{2}\hbar$ particles then come together, one from each pair, and combine to form a new particle (fig. 4). What is the spin of this new particle? Well, it must be either zero or $\hbar$, with respective probabilities $\frac{1}{4}$ and $\frac{3}{4}$ (assuming no orbital components contribute, etc.). Although you can see that there are objections even here to regarding this as giving a 'pure probability', at least the example served as a starting point. (This example was to some extent stimulated by Bohm's version of the Einstein-Rosen-Podolsky thought experiment, which it somewhat resembles.) The idea, then, is that any 'pure probability' (if such exists) ought to be something arising ultimately out of a choice between equally probable alternatives. All 'pure probabilities' ought therefore, to be rational numbers.

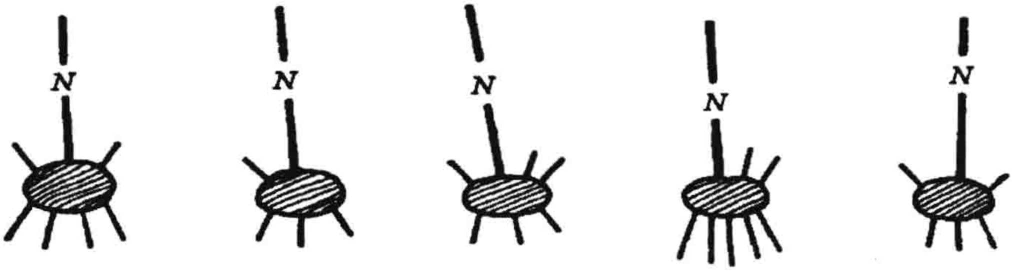


Fig. 5

But let me leave all this aside since it doesn't affect the rest of the discussion. Actually, I haven't quite finished my 'angle measuring experiment', so let me return to this.

Let us consider the following particular situation. Suppose we have a number of disconnected systems, each producing a large N-unit. There are to be absolutely no connections between them (fig. 5).

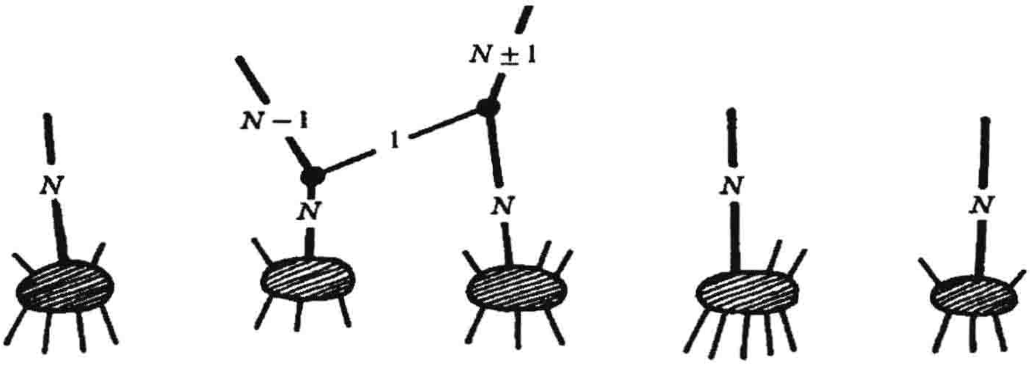


Fig. 6

Let me try to measure the 'angle' between two of them by doing one of the 'experiments' I described earlier. I detach a 1-unit from one of the N-units in such a way as to leave it as an $(N-1)$-unit. Then I reattach the 1-unit to one of the other N-units (fig. 6). According to the rules (cf. later) it will follow that the probability of the second N-unit to become an $(N \pm 1)$-unit is $\frac{1}{2}(N+1 \pm 1)/(N+1)$. These two probabilities become equal in the limit $N \to \infty$. Thus, if we are to assign an 'angle' between these units, then, for N large, this would have to be a *right-angle.* This is just using the probability blindly. I would similarly have to say, of any other pair of the N-units, that they are at right-angles. It would seem that I could put any number of N-units at right-angles to each other. In this instance I have drawn five. Does this mean that we get a five-dimensional space?—or an ∞-dimensional space?

Clearly I have not done things quite right. There are no connections between any of the N-units here, so one would like to think of the probabilities that arise out of one of these experiments as being not *just* due to the *angle* between the N-units (if they have an angle in some sense), but also due to the 'ignorance' implicit in the set-up. That is, we think of the probabilities as arising in two different ways. In the first instance, probabilities can arise in this type of experiment, if we have a *definite angle* between two spinning bodies (as we have seen). These are the genuine quantum mechanical probabilities. But, in the second instance, we may just be *uncertain* as to what the angle is between the two bodies. This lack of knowledge, concerning the history (or origins) of the two bodies, will give us a contribution to the probability value—an *ignorance* factor—which will serve to obscure the meaning of the probability in terms of angles. In the present instance, we are allowed absolutely no information concerning the interconnections between the different N-units, so the probability is

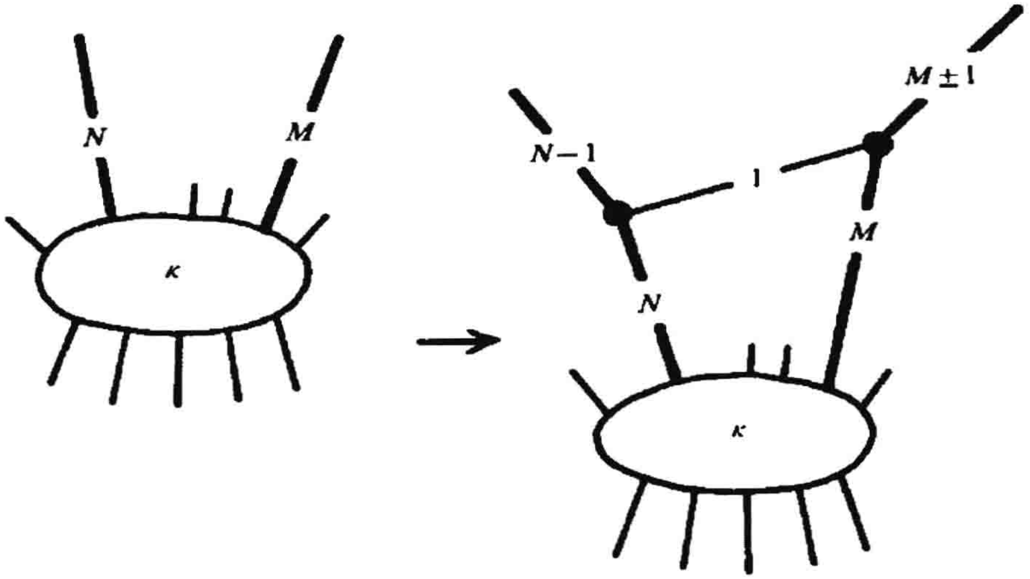


Fig. 7

not really due to ‘angle’ at all. In this extreme case, the probability is *entirely* ‘ignorance factor’. In general, the two effects will be mixed up, so we shall need a means of separating them.

Let me change the picture a bit. I’ll put in some ‘known’ connecting network (now denoted by κ) and have two large units coming out, as in fig. 7. I do one of these experiments, but then *repeat* the experiment. Suppose the N-unit is reduced to an $(N-1)$-unit and then to an $(N-2)$-unit. The M-unit becomes an $(M \pm 1)$-unit and then an $(M \pm 2)$-unit, or an M-unit again (fig. 8). The question is:

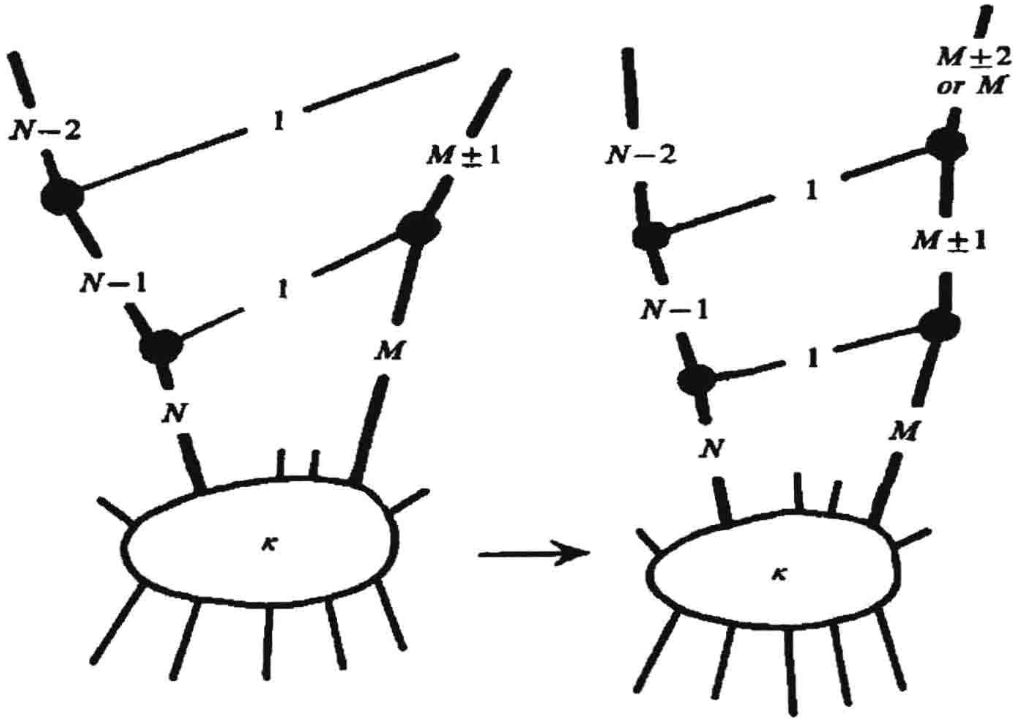


Fig. 8

is the probability of the second experiment influenced by the result of the first experiment? If this is essentially an 'ignorance' situation, where one doesn't initially know such about how the spin axes are pointing, then the *result* of the first experiment provides us with some *information* as to the relative directions of the spin axes. Therefore, the probabilities in the second experiment will be altered by the knowledge of the result of the first experiment.*

If the probabilities calculated for the second experiment are *not* substantially altered by the knowledge of the result of the first experiment, then I say that the angle between the two large units is essentially well-defined. If they *are* substantially altered by the results of the first experiment, then there is a large 'ignorance' factor involved, and the probabilities arise not just from 'angle'.

---

*It should be borne in mind that all these probability values are simply calculated here, from knowledge of the spin-networks involved. The 'experiments' are really theoretical constructions. However, it would be possible (in principle—with the usual reservations) to measure these probabilities experimentally, by simply repeating the experiment many times, each time reconstructing the spin-network afresh.

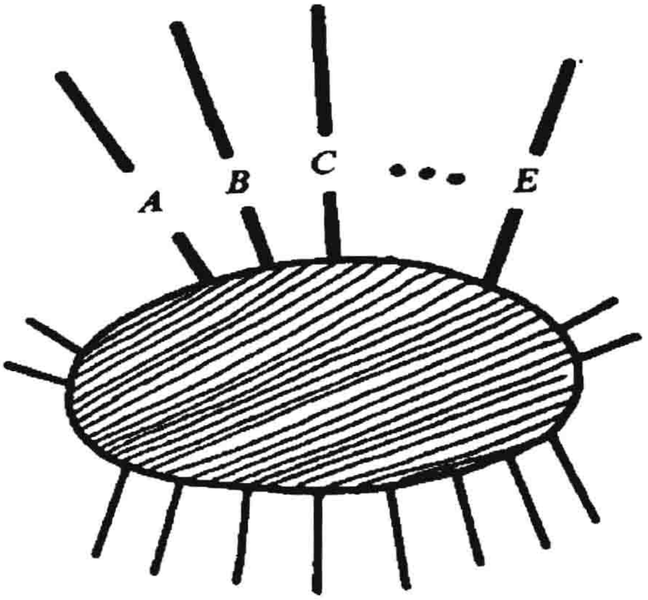


Fig. 9

Suppose, now, we have the system shown in fig. 9, which has a number of large units emerging, and suppose that it happens to be the case that the angle between any two of them is well-defined in the sense I just described. (All the numbers $A, B, \ldots$ are large compared with unity; I can do a few odd experiments which do not much change these numbers.) Then there is a theorem which can be proved to the effect that these angles are all consistent with angles between directions in Euclidean three-dimensional space.

Now, should I be in any way surprised by this result? Admittedly I should have been surprised if the method gave me any different space; but on the other hand, it is not completely clear to me that the result is something I could genuinely have inferred beforehand. Let me mention a number of curious features of the theory in this context. In the first place, suppose I set the situation up with wave functions and everything, and work according to ordinary quantum mechanical rules. I have these particles (or systems) with large angular momentum, and I finally find out that I get these angles consistent with directions in Euclidean three-dimensional space. I never, at any stage, specified that these large angular momentum systems should, in fact, correspond to bodies which do have well-defined directions (as rotation axes). There are states with large total angular momentum (e.g. $m = 0$ states) which point all over the place, not necessarily in any one direction.

I can start from some *given* Euclidean 3-space and use an ordinary Cartesian description in terms of x, y, z. I can use particles (or systems) with large total angular momentum, but which do *not* happen to give well-defined directions in the original space. Then I work out the 'angles' between them and find that these angle do not correspond to anything I can see as angles in the original description, but they are nevertheless consistent with the angle between directions in *some* abstract Euclidean three-dimensional space. I therefore take the view that the Euclidean three-dimensional space that I get out of all this, using probabilities, etc. is the *real space*, and that the original space, with its x, y, z's that I wrote down, is an irrelevant convenience, like co-ordinates in general relativity, where one writes down any co-ordinates which don't necessarily mean anything. The central idea is that *the system defines the geometry.* If you like, you can use the conventional description to fit the thing into the 'ordinary space-time' to begin with, but then the geometry you get out is not necessarily the one you put into it. So I don't know whether I should be surprised or not by the fact that I actually get the right geometry in the end.

There is a second aspect of this work that I think I regarded as slightly surprising at first. This is the fact that although no complex numbers are ever introduced into the scheme, we can still build up the full *three*-dimensional array of directions, rather than, say, a two-dimensional subset. To represent all possible directions as states of spin of a spin $\frac{1}{2}\hbar$ particle, we need to take *complex* linear combinations (in the conventional formalism). Here we only use rational numbers—and complex numbers cannot be approximated by rational numbers alone! Again, the answer seems to be that the space I end up with is not really the 'same' space as the (x, y, z)-space that I could start with—even though both are Euclidean 3-spaces.

One might ask whether corresponding rules might be invented which lead to other dimensional schemes. I don't in fact see *a priori* why one shouldn't be able to invent rules, similar to the ones I use, for spaces of other dimensionality. But I'm not quite sure how one would do this. Also, it's not obvious that the whole scheme for getting the space out in the end would still work. The rules I use are derived from the irreducible representations of SO(3). These have some rather unique features.

Now, from what I've said so far, you might wonder whether you would just scatter the numbers on the network at random. Actually, you can if you like, but unless you are a bit careful the resulting spin-network will have zero norm. And if the norm is zero, then the situation represented by the spin-network a not realizable (i.e. zero probability) according to the rules of quantum mechanics.

There are, in fact, two simple necessary requirements which must be satisfied at each vertex of a spin-network, for its norm to be non-zero. Notice first that all the spin-networks that I have explicitly drawn have the property that precisely *three* edges (i.e. units) come together at each vertex. (This isn't one of the 'requirements' I am referring to. It's just that I don't know how one would handle more general types of vertex within the scheme.) Suppose we have a vertex at which an a-unit, a b-unit and a c-unit come together (fig. 10).

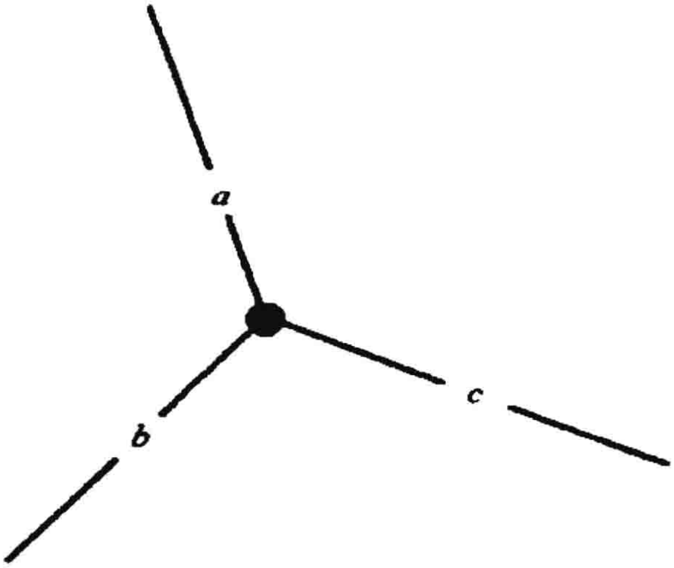


Fig. 10

Then for a spin-network containing the vertex to have a non-zero norm, it is necessary that the triangle inequality hold:

$$a+b+c \geq 2\max(a,b,c) \quad ;$$

and furthermore that there be conservation of fermion number (mod 2):

$$a+b+c \quad \textit{is even.}$$

These are, of course, properties that one would want to hold in real physical processes, with the interpretations that I have given to the spin-networks.

But even if these requirements hold at every vertex, the spin-network may still have zero norm. For example, each of the two types of spin-network shown in fig. 11 has zero norm, where $n \neq 0$ in the first case and $n \neq m$ the second. In each case, the shaded portion represents some spin-network with no other free ends. In fact, the first is effectively a special case of the second, with $m = 0$. This is because any 0-unit can be omitted from a spin-network (if we also suitably delete the relevant vertices) without changing the norm. We may interpret the

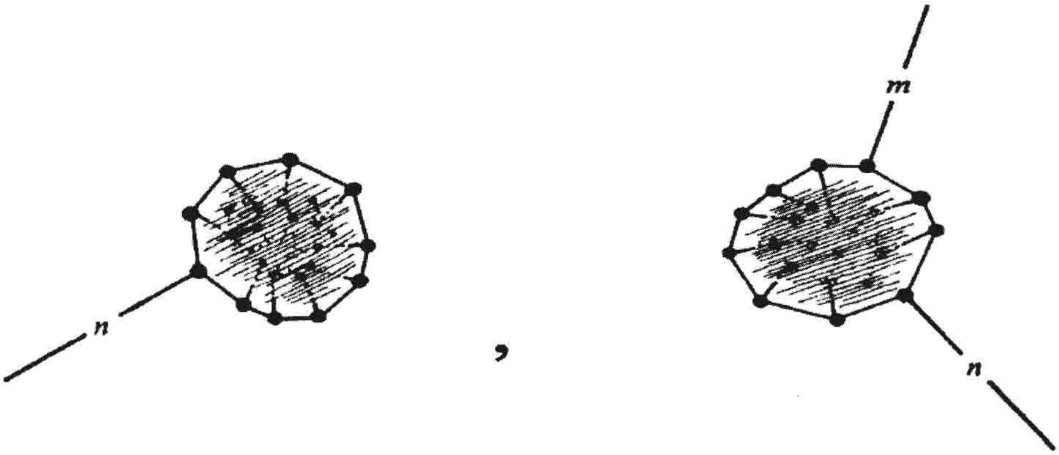


Fig. 11

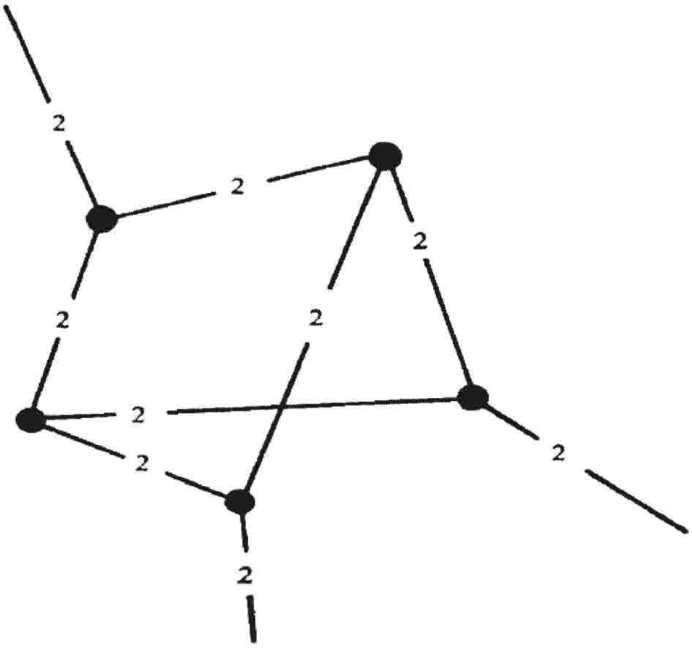


Fig. 12

vanishing of the norm whenever $n \neq m$ in the second case as an expression of *conservation of total angular momentum.*

In addition to these cases, there are many particular spin-networks which turn out to have zero norm. One example is shown in fig. 12. But so far I have only been giving particular cases. Let us now pass to the general rule.

I shall give the definition of the norm in terms of a closely related concept, namely, what I shall call the *value* of a closed oriented spin-network. I call a spin-network *closed* if it has no free ends (e.g. analogous to a disconnected vacuum process). A spin-network which is not closed will not be assigned a value. The definition of *orientation* for a spin-network is a little difficult to give concisely. Any spin-network can be assigned two alternative orientations. Fixing the orientation of a closed spin-network will serve to define the *sign* of its value (which can be positive or negative). Roughly speaking the orientation assigns a cyclic order to the three units attached to each vertex—but if we reverse the cyclic order at any *even* number of vertices this is to leave the orientation unchanged. The orientation will change, on the other hand, if the cyclic order is reversed at an *odd* number of vertices.

I shall adopt the convention, when drawing spin-networks, that the orientation is to be fixed by the way that the spin-network is depicted on the plane. At each vertex we specify '*counter-clockwise*' as the cyclic order for the three units attached to the vertex. This defines the spin-network's orientation. The diagrams in fig. 13 illustrate an example of a closed spin-network with its two possible orientations.

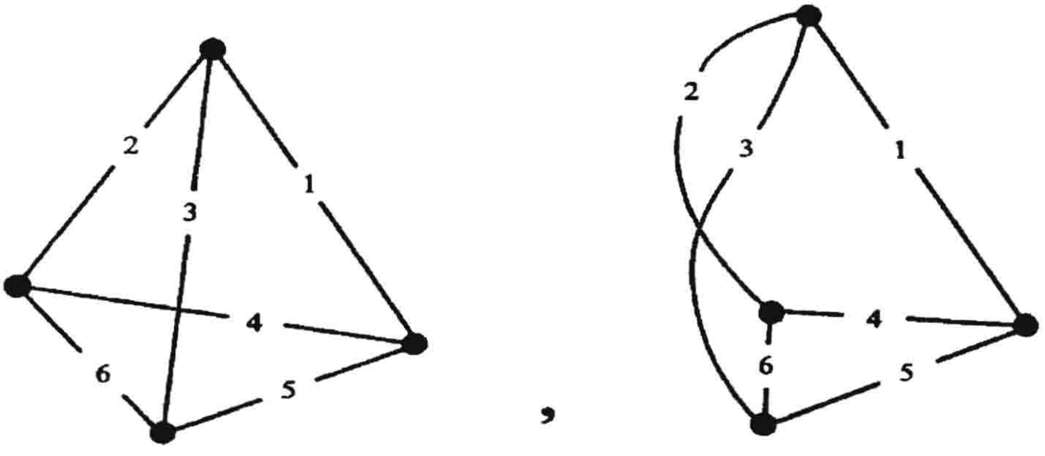


Fig. 13

It will also be convenient to use the representation of a spin-network as a drawing on a plane, in order to keep track of signs properly when defining the value. This may have the effect of making the definition *seem* less 'combinatorial' than it really is. Of course, the definition could be reformulated without the use of such a drawing if desired. Consider, then, a closed spin-network α depicted as a

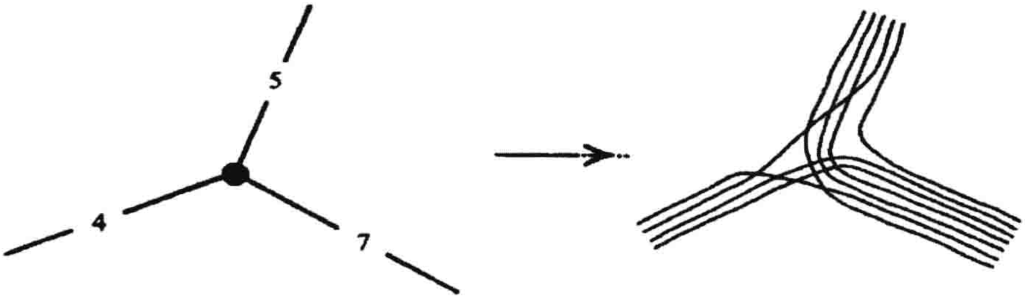


Fig. 14

drawing on a plane. Now, imagine each n-unit to be replaced in the drawing by n parallel strands. At each vertex, the strand ends must be connected together in pairs, but no two strands associated with the same n-unit are to be connected together. Let us call such a connection scheme a *vertex connection.* One such vertex connection is illustrated in fig. 14, while fig. 15 shows a non-allowable connection,

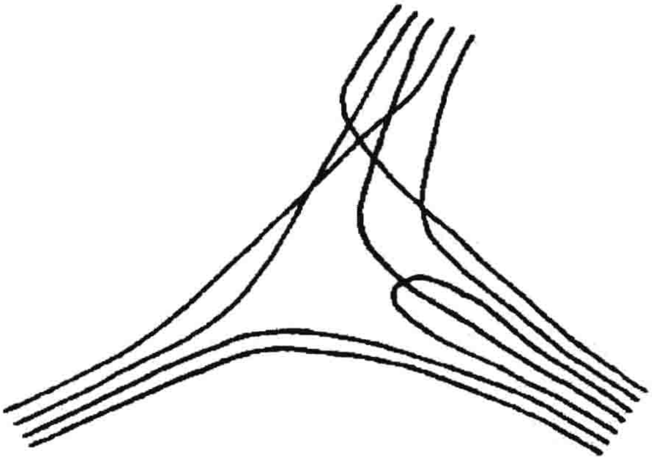


Fig. 15

since two strands of the 7-unit are connected to one another. The *sign* of a vertex connection is defined most simply as $(-1)^x$ where x is the number of intersection points between different strands at the vertex, as drawn on the plane. (These intersection points must be counted correctly if more than two strands cross at a point, or if two strands touch: and ignored if a strand crosses itself. It is simplest on the other hand, just to avoid such features by drawing the strands in general position and not allowing any strand connection to cross itself.) The sign of a vertex connection, in fact, does not depend on the details of how it is drawn, but only on the pairing off of the strands. The allowable vertex connection depicted above has -1 as its sign, since there are thirteen crossing points.

When the vertex connections have been completed at every vertex of a closed spin-network, then we shall have a number of *closed loops*, with no open-ended strands remaining. Consider, now, every possible way of allowably completing the vertex connection for the spin-network α. We form the expression

$$\text{value of } \alpha = \frac{\sum \pm(-2)^c}{\prod n!}$$

where the summation extends over all possible completed allowable connection schemes, where the '±' stands for the product of the signs of all the vertex connections, where c is the number of closed loops resulting from the vertex connections and where the product in the denominator ranges over all the units of the spin-network, n being the spin-number of the unit. The value of any closed spin-network always turns out to be an integer.

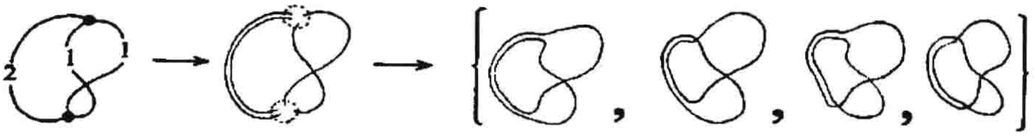


Fig. 16

$$\begin{aligned} \rightarrow \text{value} &= \frac{1}{2!\,1!\,1!}\{+(-2)-(-2)^2-(-2)^2+(-2)\} \\ &= -6. \end{aligned}$$

Let us consider a simple example, given in fig. 16. Note that the 'accidental' intersection, arising from the crossing of the two 1-units in the first drawing of the spin-network, does not contribute to the sign of the terms in the sum. Only the intersections at the vertex connections count.

The definition of the value of a closed oriented spin-network that I have just given is perhaps the simplest to state, but it is by no means the most useful to use in actual calculations. When the spin-networks become even slightly more complicated than the simple one evaluated above, the detailed calculations can become very unwieldy. A more useful procedure is to employ certain reduction formulae which can be used to express complicated networks in terms of simpler ones.[†] For this purpose, it will be necessary to introduce a slight variation of the spin-network theme; I shall consider the related concept of a *strand-network*.

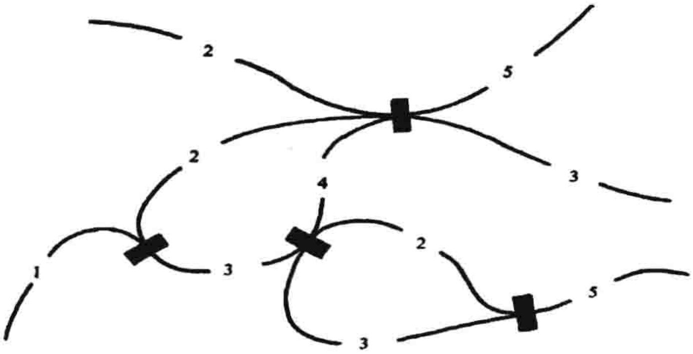


Fig. 17

A strand-network is a series of connections relating objects (which I shall still refer to as n-units) an example of which is depicted in fig. 17. The units are 'tied together' at various places, as indicated by the thick bar. Any spin-network can be translated into strand-network terms, by replacing each vertex according to the scheme shown in fig. 18. I thus introduce three more ('virtual') units at each vertex. A strand-network is *closed* if it has no free ends. Any closed (oriented) strand-network will have a *value* which is an integer (positive, negative or zero). This value will be chosen to agree with that defined for a spin-network, in the case of closed strand-networks obtained by means of the above replacement. Generally, to obtain the value of s closed strand-network β we employ the same formula as before:

$$\text{value of } \beta = \frac{\sum \pm(-2)^c}{\prod n!}$$

[†] Diagrams closely related to spin-networks were introduced by Ord-Smith and Edmonds for the graphical treatment of quantum mechanical angular momentum. (See reference (1) for a detailed discussion.)

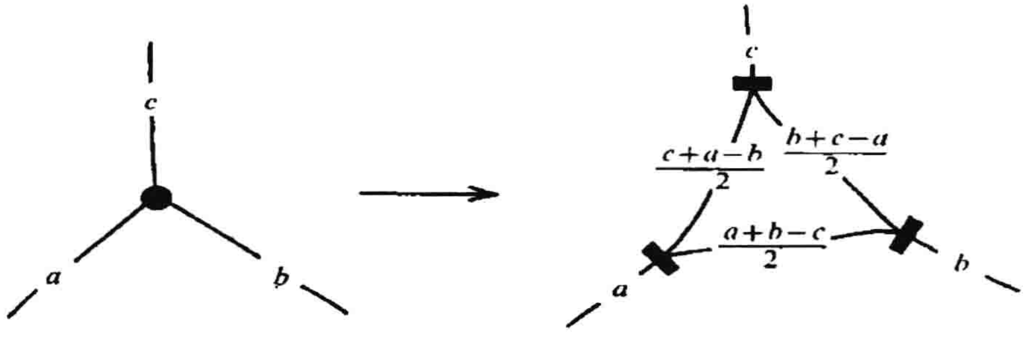


Fig. 18

where now the '±' sign refers to the product of all the signs of all the permutations involved in each strand connection at which the strands come together. For example, one possible connection scheme for

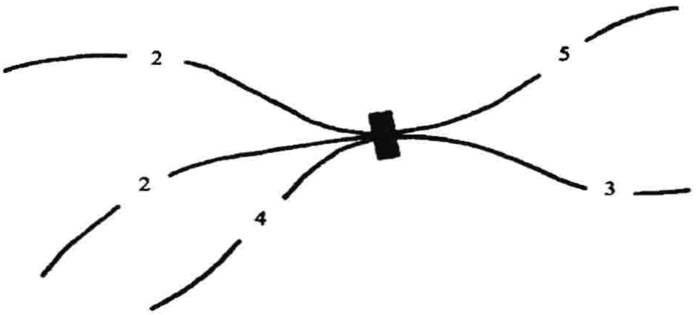


Fig. 19

'strand vertex' of fig. 19, would be that shown in fig. 20. This connection scheme would contribute a minus sign, since an odd permutation is involved. (There are nine crossing points—this is essentially

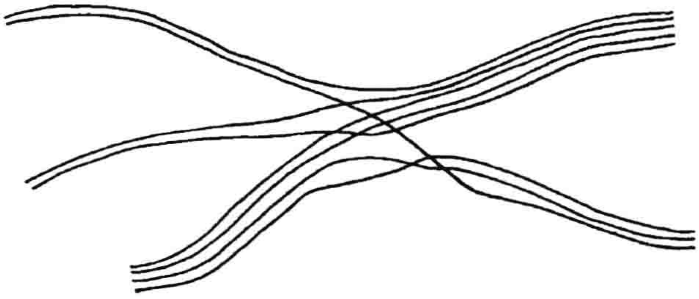


Fig. 20

the 'Aitken diagram' method of determining the sign of a permutation.) Notice that for a connection scheme to be possible at all, we require that the total of the spin-numbers entering at one side must equal the total of the spin-numbers leaving at the other. This *one* requirement now takes the place of the 'triangle inequality' and 'fermion conservation' that we had earlier.

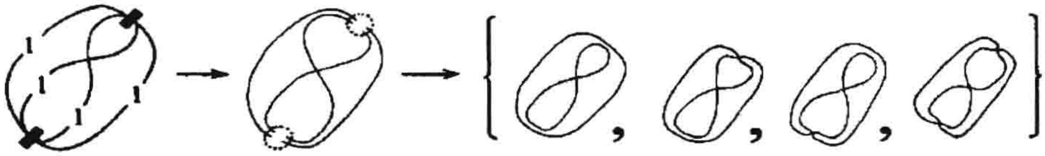


Fig. 21

$$\begin{aligned} \therefore \text{ value } &= \frac{1}{1!\,1!\,1!\,1!}\{+(-2)^2-(-2)-(-2)+(-2)\} \\ &= 2. \end{aligned}$$

Let us evaluate the simple closed strand-network of fig. 21 as an example. Again there is an 'accidental' intersection depicted (where two 1-units cross) which does not contribute to the sign of the terms in the sum.

Let me list a number of relations and reduction formulae which are useful in evaluating strand-networks (fig. 22). (I am not going to prove anything here, but most of the relations are not hard to verify.) These relations may be substituted into any closed strand-network and a valid relation between values is obtained. Finally, let me make the remark that the value is *multiplicative*, that is to say, the value of the union of two *disjoint* strand-networks or spin-networks is equal to the product of their individual values.

I now come to the definition of the *norm* of a spin-network. A strand-network will likewise have a norm. This is simply obtained by drawing *two* copies of the spin-network (or strand-network), joining together the corresponding free end units to make a closed network, and then taking the modulus of the value of this resulting closed network. As an example, the norm of the spin-network consisting of a single vertex is found in fig. 23. An even simpler example, depicted in fig. 24, is the norm of a single isolated n-unit.

Finally, let me describe how the norm may be used in the calculation of *probabilities* for spin-numbers, in the type of 'experiment' that we have been considering. (Again I shall give no proofs.) Suppose we start with a spin-network α, with an a-unit and a b-unit among its

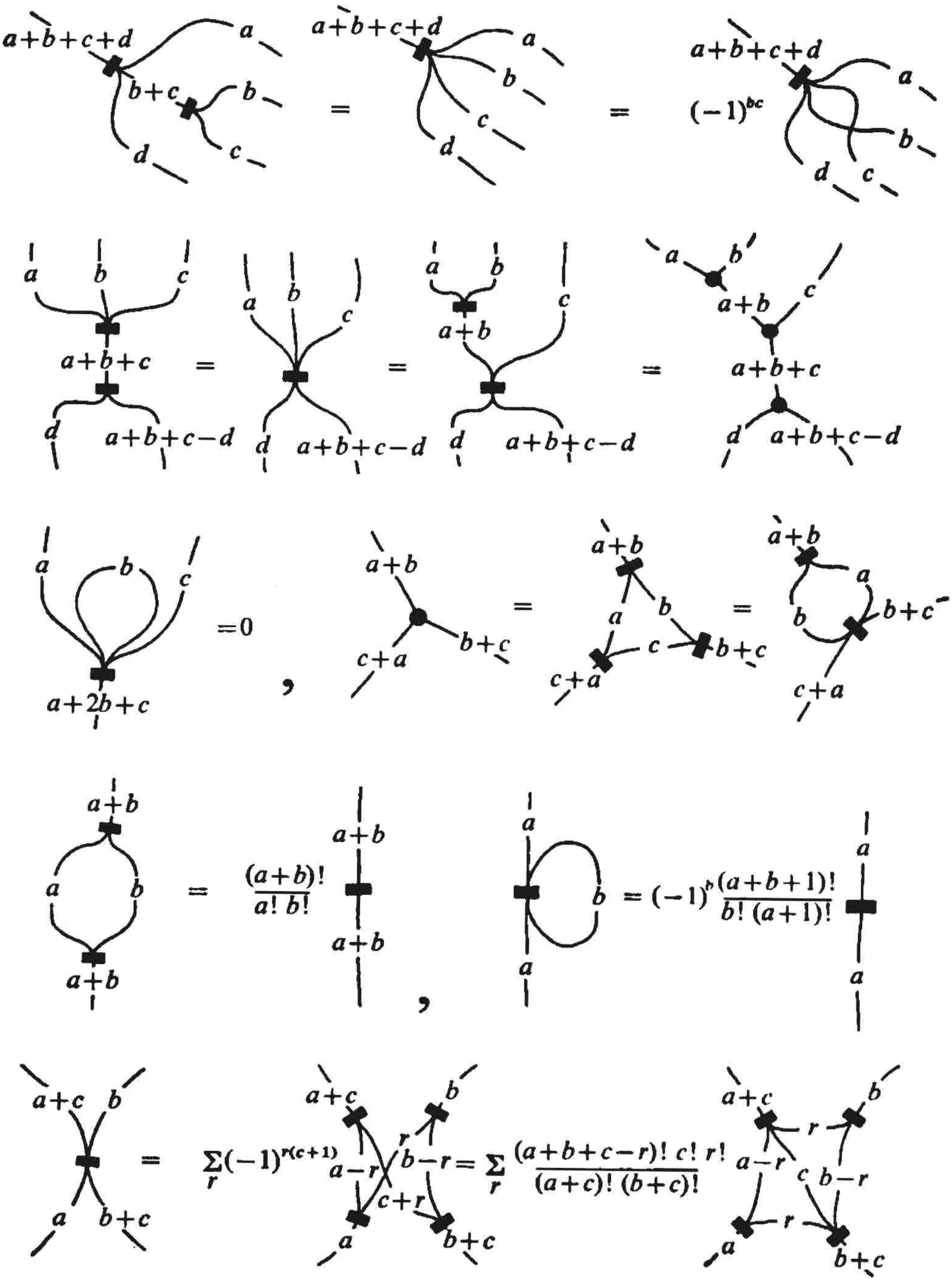

a+b+c+d
a
b+c
b
c
d
=
a+b+c+d
a
b
c
d
=
a+b+c+d
(−1)^{bc}
a
b
d
c
a
b
c
a+b+c
d
a+b+c−d
=
a
b
c
d
a+b+c−d
=
a
b
a+b
c
d
a+b+c−d
=
a
b
a+b
c
a+b+c
d
a+b+c−d
a
b
c
a+2b+c
=0
,
a+b
b+c
c+a
=
a+b
a
b
c
c+a
b+c
=
a+b
a
b
b+c
c+a
a+b
a
b
a+b
=
(a+b)!/(a! b!)
a+b
a+b
,
a
b
a
= (−1)^b (a+b+1)!/(b! (a+1)!)
a
a
a+c
b
a
b+c
=
Σ_r (−1)^{r(c+1)}
a+c
b
r
a−r
b−r
c+r
a
b+c
= Σ_r (a+b+c−r)! c! r! / ((a+c)! (b+c)!)
a+c
b
r
a−r
c
b−r
r
a
b+c

Fig. 22

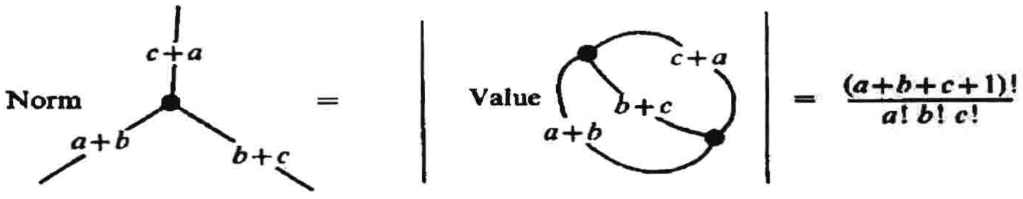

Norm
c+a
a+b
b+c
=
Value
c+a
b+c
a+b
=
(a+b+c+1)!
a! b! c!

Fig. 23

free ends (fig. 25). Suppose the a-unit and the b-unit come together to form an x-unit, the resulting spin-network being denoted by β (see fig. 26). We wish to know (given α) what are the various probabilities for the different possible values of the spin-number x. Let γ denote the spin-network representing the coming together of the a-unit and

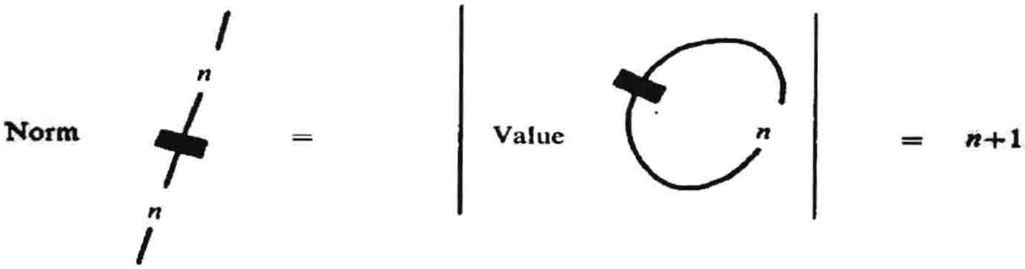

Norm
n
n
=
Value
n
=
n+1

Fig. 24

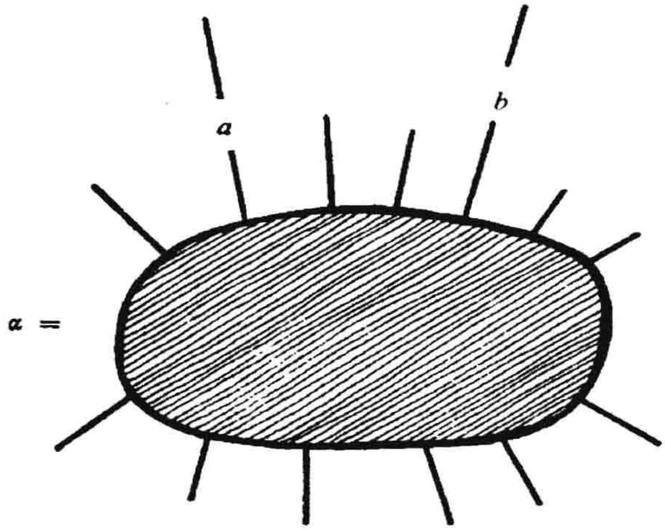


Fig. 25

the b-unit to form the x-unit. Let ξ denote the spin-network consisting of the x-unit alone. These are illustrated in fig. 27. Then the required probability for the resulting spin-number to be x is

$$\text{probability} = \frac{\text{norm}\beta\,\text{norm}\xi}{\text{norm}\alpha\,\text{norm}\gamma}\,.$$

Using the explicit expressions for norm γ and norm ξ that were just given as examples (using a slightly different notation), we can rewrite this as

$$\text{probability} = \frac{\text{norm}\beta\,(x+1)\,\{\tfrac{1}{2}(a+b+x)+1\}!}{\text{norm}\alpha\,\{\tfrac{1}{2}(a+b-x)\}!\,\{\tfrac{1}{2}(b+x-a)\}!\,\{\tfrac{1}{2}(x+a-b)\}!}$$

From the combinatorial nature of the definition of norm, it is clear that these probabilities must all be rational numbers. And with the interpretation of 'angle' that I have given, the three-dimensional Euclidean nature of the 'directions in space' that are obtained, is a *consequence* of these combinatorial probabilities.

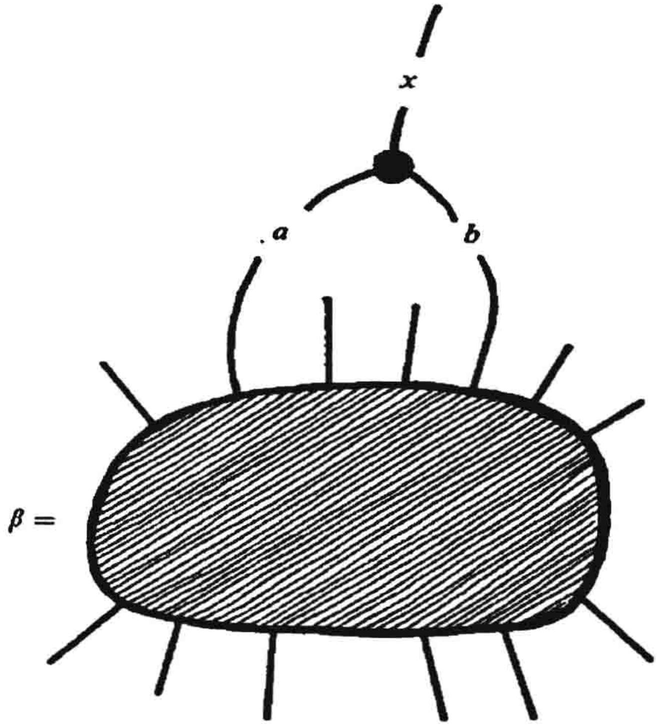


Fig. 26

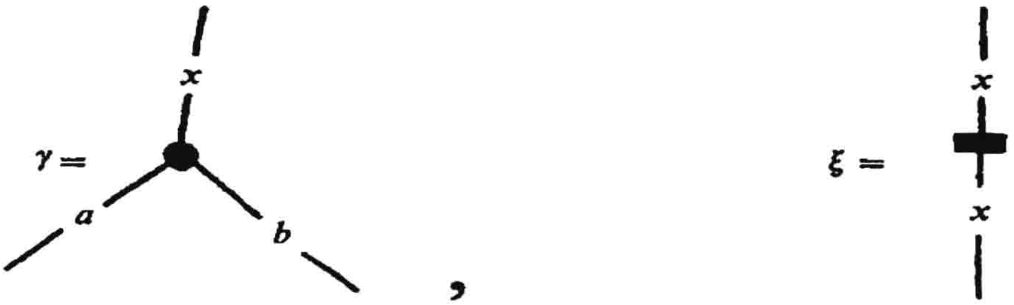


Fig. 27

I should emphasize again that the space that I get out in the end is the one *defined* by the system itself and is not really the same space as the one that might have been introduced at the start if a conventional formalism had been used. Thus, although undoubtedly the reason that we end up with directions in a Euclidean three-dimensional space is intimately *related* to the fact that we start with representations of the rotation group SO(3), the precise logical connections are not at all clear to me. When I come to consider the generalization of all this to a relativistic scheme in a moment, this question will again present itself. I shall also need to consider the spatial locations of objects, not just their orientations. My model works with objects and the interrelations between objects. An object is thus 'located', either directionally or positionally in terms of its relations with other objects. One doesn't really need a space to begin with. The notion of space comes out as a *convenience* at the end.

Essentially, I have so far been using a non-relativistic scheme. The angular momentum is not relativistic angular momentum. From a four-dimensional viewpoint, the directions I get are those orthogonal to a given timelike direction, i.e. directions in three-dimensional space. All the particles are going along in this same timelike direction. Perhaps they can knock each other a little bit, but they are not really moving very much. They just transfer angular momentum backwards and forwards. All the particles are, strictly speaking in the same place, not moving relative to one another. Consequently, one does not have any problem of mixing between orbital and spin angular momentum. Once I allow orbital contributions to come in, then I must drastically change the scheme, since now not only is the question of 'direction' and 'angle' involved, but so also is 'position' and 'distance'. Thus, if one thinks of *real* particles moving relative to each other, then there is the problem not only of doing things relativistically, but also of bringing in actual displacements between particles. Consider two particles in relative motion. Suppose they come together and combine to form a system with a well-defined total spin. Then to obtain the spin of the combined system, we cannot just add up the individual spins because we have to bring in the orbital component. There is a mixture of the actual displacements in space with the angular momentum concept. So I spent a long time thinking how one should combine rotations and displacements together into an appropriate relativistic scheme. Eventually I was led to consider a certain algebra for space-time which treats linear displacements on the same footing as it treats rotations. Thus, linear momentum is treated on a similar footing to angular momentum.

Now you might raise the objection that linear momentum has a continuous spectrum, while it is only for angular momentum that one has discreteness. This is a problem of some significance to us. My answer to it is roughly the following: each particle has its own discrete spectrum for its angular momentum. When two particles are considered together as a unit, then again there is a discrete spectrum for the combined system. The way these 'spins' add up implicitly brings in the relative motion between the two particles. So the momentum is brought in through the back door, in a sense, for one could be always talking in terms of 'bound systems'.

I consider momentum states as being linear combinations of angular momentum states. There is indeed a problem to see how this continuous momentum should be built up from something discrete, but, in principle, there is nothing against it. In effect, the idea is that the momentum should be brought in indirectly. I would propose that, in a sense, there should *not* be a well-defined distinction between momentum and angular momentum—except in the limit. Individual particles and simple systems would not really 'know' what momentum is. Like the idea of 'direction' that I considered earlier, it would be only in the limit of large systems that the concept of momentum really attains a well-defined meaning. Smaller systems might retain a combined concept of momentum and angular momentum, but these things would only sort themselves out properly in the limit.

The algebra I have used to treat linear displacements and rotations together, or linear and angular momentum together, I call the algebra of *twistors*. I have used the term 'twistor' to denote a 'spinor' for the six-dimensional $(+ + - - - -)$ pseudo-orthogonal group O(2,4). The twistor group is the $(+ + - -)$ pseudo-unitary group SU(2,2), which is locally isomorphic with O(2,4). In turn, O(2,4) is locally isomorphic with the fifteen-parameter (local) conformal group of space-time. Under a conformal transformation of space-time, the twistors will transform (linearly) according to a representation of the group SU(2,2).

The basic twistor is a four-complex-dimensional object. We can thus describe it by means of four complex components Z^α:

$$(Z^\alpha) = (Z^0,\, Z^1,\, Z^2,\, Z^3)\,.$$

The complex conjugate of the twistor Z^α is an object $\bar{Z}_\alpha$ whose components

$$(\bar{Z}_\alpha) = (\bar{Z}_0,\, \bar{Z}_1,\, \bar{Z}_2,\, \bar{Z}_3)$$

are given (according to a convenient co-ordinate system) by

$$\bar{Z}_0 = \overline{(Z^2)}, \quad \bar{Z}_1 = \overline{(Z^3)}, \quad \bar{Z}_2 = \overline{(Z^0)}, \quad \bar{Z}_3 = \overline{(Z^1)}.$$

This implies that the Hermitian form $Z^\alpha \bar{Z}_\alpha$ (summation convention assumed) has signature $(+ + - -)$, which is required in order to give SU(2,2). (I have already described these objects(2) and their geometrical significance in Minkowski space-time, and a later paper(3) goes into some further developments, including some of the quite surprising aspects of the theory which arise when one starts to describe physical fields in twistor terms.)

When $Z^\alpha \bar{Z}_\alpha = 0$ I call Z^α a *null* twistor. A null twistor Z^α has a very direct geometrical interpretation in space-time terms. In fact, Z^α defines a *null straight line*, which we can think of as the world line of a zero rest-mass particle. (An important aspect of twistor theory is that *zero rest-mass* is to be regarded as more fundamental than a finite rest-mass. Finite mass particles are viewed as composite systems, the mass arising from interactions.) The twistor λZ^α, for any non-zero complex number λ, defines the same null line as does Z^α. But we can distinguish Z^α from λZ^α by assigning to the particle a 4-momentum (in its direction of motion) and, in addition, a sort of 'polarization' direction (both constant along the world line of the particle). When Z^α is replaced by rZ^α (r real) the momentum is multiplied by r^2. When Z^α is replaced by $e^{i\theta} Z^\alpha$ (θ real) the 'polarization' plane is rotated through an angle 2θ. If Y^α is another null twistor, the condition for the null lines represented by Y^α and Z^α to *meet* is

$$Y^\alpha \bar{Z}_\alpha = 0,$$

i.e. this is the condition for the two particles to 'collide'. (To be strictly accurate we have to include the possibility that they may 'meet at infinity'.

In addition, some of the null twistors describe 'null lines at infinity' rather than actual null lines.)

The *non-null* twistors are divided into two classes according as $Z^\alpha \bar{Z}_\alpha$, is positive or negative. If $Z^\alpha \bar{Z}_\alpha > 0$, I call Z^α right-handed; if $Z^\alpha \bar{Z}_\alpha < 0$, left-handed. In Minkowski space-time, one can give an interpretation of a non-null twistor in terms of a twisting system of null lines. The helicity of the twist is defined by the sign of $Z^\alpha \bar{Z}_\alpha$. In more physical terms, the twistor Z^α (up the phase) describes the momentum and angular momentum structure of a zero rest-mass particle.[‡] We can make the interpretation that $Z^\alpha \bar{Z}_\alpha$, is (twice) the *intrinsic spin* of the particle, measured in suitable units, with a *sign* defining the helicity. If $Z^\alpha \bar{Z}_\alpha \neq 0$, then it is not possible actually to localize the particle as a null straight line. Only if $Z^\alpha \bar{Z}_\alpha = 0$ do we get a uniquely defined null line which we can think of as the world line of the particle; otherwise the particle to some extent spreads itself throughout space.

The twistor co-ordinates Z^0, Z^1, Z^2, Z^3, together with their complex conjugates $\bar{Z}_0$, $\bar{Z}_1$, $\bar{Z}_2$, $\bar{Z}_3$, can be used in place of the usual x, y, z, t and their canonical conjugates p_x, p_y, p_z, E. In fact, anything that can be written in normal Minkowski space terms can be rewritten in terms of twistors. However, in principle, the twistor expressions for even quite simple physical processes may turn out to be very complicated. But in fact it emerges that the basic elementary processes that one requires, can actually be expressed very simply if one goes about it in the right way. Analytic (holomorphic) functions in the Z^α variables play a key role. So does contour integration.

We can regard the Z^α as quantum operators under suitable circumstances. Then $\bar{Z}_\alpha$, can be regarded as the canonical conjugate of Z^α. (I shall go into the reasons for this a little more later.) We have commutation rules

$$Z^\alpha \bar{Z}_\beta - \bar{Z}_\beta Z^\alpha = \delta^\alpha{}_\beta \hbar$$

[‡] Using a convenient co-ordinate system, we can relate the momentum P_α and the angular momentum $M^{ab}(= -M^{ba})$ to the twistor variables Z^α, $\bar{Z}_\alpha$ a by:

$$P_0 + P_1 = 2^{\frac{1}{2}} Z^0 \bar{Z}_2, \quad P_0 - P_1 = 2^{\frac{1}{2}} Z^0 \bar{Z}_3, \quad P_2 + i P_3 = 2^{\frac{1}{2}} Z^1 \bar{Z}_2$$
$$P_2 - i P_3 = 2^{\frac{1}{2}} Z^0 \bar{Z}_3, \quad M^{23} + i M_{01} = Z^1 \bar{Z}_1 - Z^0 \bar{Z}_0, \quad M^{23} - i M_{01} = Z^3 \bar{Z}_3 - Z^2 \bar{Z}_2$$
$$M^{13} + M^{03} + i M_{02} + i M_{12} = 2 Z^2 \bar{Z}_2, \quad M^{13} - M^{03} + i M_{02} - i M_{12} = 2 Z^3 \bar{Z}_2$$
$$M^{13} + M^{03} - i M_{02} - i M_{12} = 2 Z^1 \bar{Z}_0, \quad M^{13} - M^{03} - i M_{02} + i M_{12} = 2 Z^0 \bar{Z}_1$$

Then, since Z^α and $\bar{Z}_\alpha$ do not commute, we must re-interpret the expression for the *spin-helicity* $\frac{1}{2}n$ as the *symmetrized* quantity,

$$\frac{1}{4}(Z^\alpha\,\bar{Z}_\alpha - \bar{Z}_\alpha\,Z^\alpha) = \frac{1}{2}n\hbar$$

Only zero rest-mass states can be eigenstates of this operator. The eigenvalues of n are $\ldots -2, -1, 0, 1, 2, \ldots$ The operators for the ten components of momentum and angular momentum (together with those for the five extra components arising from the conformal invariance of zero rest-mass fields) are

$$Z^\alpha\,\bar{Z}_\beta - \frac{1}{4}\delta^\alpha{}_\beta\,Z^\gamma\,\bar{Z}_\gamma$$

in twistor notation. The usual commutation rules for momentum and angular momentum are then a consequence of the twistor commutation rules.

One idea would now be to use this fact and simply let the twistors take the place of the *two-component spinors* that lay 'behind the scenes' in my previous non-relativistic approach, and then to attempt to build a concept of a four-dimensional space-time from whatever graphical algebra arises from the twistor rules. I have not attempted to do quite this, as yet, since I am not sure that it is exactly the right thing to do. There are certain other aspects of twistor theory which should really be taken into account first.

Let me mention one particular point. It is a rather remarkable one. If the twistor approach is going to have any fundamental significance in physical theory, then it ought, in principle at least, to be possible to carry the formalism over and apply it to a *curved* space-time, rather then just a flat space-time. These objects, as originally defined, are very much tied up with the Minkowski flat space-time concept. How can we carry them over into a curved space-time? Actually, a twistor for which $Z^\alpha\,\bar{Z}_\alpha = 0$ carries over very well. Its interpretation is now simply as a *null geodesic* (i.e. world line of a freely moving massless particle) with a momentum (pointing along the world line) and a 'polarization' direction (both covariantly constant along the world line). On the other hand, it does not seem to be possible to interpret a *non-null* twistor, in a general curved space-time, in precise classical space-time terms. Nevertheless it turns out to be convenient to *postulate* the existence of these non-null twistors—as objects with no classical realization in curved space-time terms. (In a sense, twistors are more appropriate to the treatment of *quantized* gravitation§ than of classical general relativity.)

---

§Since this lecture was delivered, there have been some developments in twistor theory

Let us concentrate attention, for the moment, on the *null* twistors only so that we can consider purely geometrical questions. We are interested in properties of null geodesics which refer to each geodesic as a *whole* and not to the neighbourhood of some point on a geodesic. Consider, for example, the condition of orthogonality between twistors. We have seen that in flat space-time, the condition of orthogonality $Y^\alpha \bar{Z}_\alpha = 0$ between two twistors Y^α, Z^α corresponds to the *meeting* of the corresponding null lines. In curved space-time this is not really satisfactory, because although I can tell whether two null geodesics are going to meet if I look in the neighbourhood of the intersection point, if I look somewhere else, I can't see whether or not they will meet, because the curvature may have bent them away from each other. So, in fact, the orthogonality property is not something which is preserved, as an invariant concept, when one turns to curved space-time. On the other hand, certain things *are* preserved; and, somewhat surprisingly, they correspond to assigning a *symplectic* structure to the twistor space.

This symplectic structure is expressed (in appropriate co-ordinates) as the invariance of the 2-form

$$dZ^\alpha \wedge d\bar{Z}_\alpha$$

(using Cartan notation). Strictly speaking, this requires the postulated non-null twistors, in addition to the null ones. The null twistors only form a seven-real-dimensional manifold, whereas a symplectic manifold must be even-dimensional. The null twistor manifold must be embedded in the eight-real-dimensional manifold consisting of *all* twistors. The structure of the null twistor manifold is that induced by the embedding in this eight-dimensional symplectic manifold. In addition to the symplectic structure (and closely related to it), the expressions

$$Z^\alpha \bar{Z}_\alpha, \quad Z^\alpha d\bar{Z}_\alpha, \quad Z^\alpha \frac{\partial}{\partial Z^\alpha}$$

---

of relevance to this discussion. It seems to be possible to express quantized gravitational theory in twistor form. By means of $3k$-dimensional contour integrals in spaces of many twistor variables, one can apparently calculate scattering amplitudes for processes involving gravitons, photons and other particles. Diagrams arise which can be used to replace the spin-networks of the formalism described here. It is not impossible that the calculations can be reduced to a set of comparatively simple combinatorial rules, but it is unclear, as yet, whether this is so. The work is still very much at a preliminary stage of development and many queries remain unanswered.

are also invariant. All these invariant quantities can be interpreted, to some extent, in terms of the geometrical properties of null geodesics. But it will not be worthwhile for me to go into all this here.

The invariance of the symplectic structure of the twistor space for curved space-time can be re-expressed as the invariance of the *Poisson brackets*

$$[\psi, \chi] = i\,\frac{\partial\psi}{\partial Z^\alpha}\frac{\partial\chi}{\partial\bar{Z}_\alpha} - i\,\frac{\partial\chi}{\partial Z^\alpha}\frac{\partial\psi}{\partial\bar{Z}_\alpha}$$

This strongly suggests that in the passage to quantum theory, $\bar{Z}_\alpha$ should be regarded as the conjugate variable to Z^α. Thus, we are led to the commutation rules I mentioned earlier, relating quantum operators Z^α and $\bar{Z}_\alpha$. These commutation rules in turn give us the commutation relations for momentum and angular momentum, as I indicated before. This suggests that there may possibly be some deep connection between these commutation rules (or perhaps some slight modifications of them) and the curvature of space-time.

The picture that one gets is that in some sense the curvature of space-time is to do with canonical transformations between the twistor variables Z^α, $\bar{Z}_\alpha$. Suppose we start in some region of space-time where things are essentially flat. Then we can interpret Z^α and $\bar{Z}_\alpha$ in a straightforward way in terms of geometry and angular momentum, etc. Suppose we then pass through a region of curvature to another region where things are again essentially flat. We then find that our interpretations have undergone a 'shift' corresponding to a canonical transformation between Z^α and $\bar{Z}_\alpha$. In effect the 'twistor position' (i.e. Z^α) and 'twistor momentum' (i.e. $\bar{Z}_\alpha$) have got mixed up. Somehow it is this mixing up of the 'twistor position' and 'twistor momentum' which corresponds to what we see as space-time curvature.

Going back to my original combinatorial approach for building space up from angular momentum, we can ask now whether such a combinatorial scheme could be applied to the twistors. Might it be that, instead of ending up with a flat space, we could end up with a curved space-time? Even if I start with the commutation rules appropriate just to the Poincaré group, or perhaps the conformal group, it is obvious that I must end up with essentially the same space that I 'start' with? One has to define the things with which one builds up geometry (e.g. points, angle, etc.), in *terms* of the physical objects under consideration. It is not at all clear to me that the geometry that is built up in one region will not be 'shifted' with respect to the geometry built up in some other region. Is it then not possible that a space-time possessing curvature might be the result? That is really the final point I wanted to make.

REFERENCES

(1) Yutsis, A. P., Levinson, I. B. and Vanagas, V. V. *Mathematical apparatus of the theory of angular momentum* (Jerusalem: I.P.S.T., 1962).
(2) Penrose, R. *J. Math. Phys.* (1967) **8**, 345.
(3) Penrose, R. *Int. Jl Theor. Phys.* (1968) **1**, 61.

47

EXTRACTION OF ROTATIONAL ENERGY FROM A BLACK HOLE

Nature. Physical Science **229** (1971), 177. With R. Floyd

Although the specific method of extracting rotational energy from a black hole by means of a particle bifurcation process (now frequently referred to as the 'Penrose process') appeared in a footnote in the article of Chapter 40, p. [272], it was the specific calculation demonstrated in the present paper that showed this procedure to be actually possible. Reva Williams and others have argued that actual energy emissions from black holes (quasars, radio galaxies, etc.) might explicitly use this process, but the general consensus seems to be that magnetic fields play a more important role. However, whatever procedure Nature uses for the actual extraction of rotational energy from a black hole, it must depend on the Killing-vector property described here and in Chapter 40 (existence of an 'ergosphere'). Although the area-increase principle for black holes is tentatively referred to at the end of the present article, I was rather too cautious—especially in view of later developments—in stating the generality of this principle as was known to me at the time. This article suffers from a notational curiosity in the printing that makes it hard to read, where the numbers in parentheses above various symbols are far too large.

Extraction of Rotational Energy from a Black Hole

THERE has been considerable interest recently in the question of the gravitational collapse of a massive body and of the possible astrophysical consequences of the existence of the "black hole" which general relativity predicts should sometimes be the result of such a collapse. In particular, the question has arisen whether the mass–energy content of a black hole could, under suitable circumstances, be a source of available energy. We now consider the extraction of rotational energy from a black hole, not least because the rotational energy (defined appropriately) of a black hole should, in general, be comparable with its total mass–energy[1].

The extraction of rotational energy from neutron stars is important in the theory of pulsars (in which case the rotational energy is of the general order of only 10^{-5} of the mass–energy of the object), where the coupling between angular momentum of the star and the surrounding medium is presumed to be achieved by means of magnetic field lines in plasma. It is exceedingly unlikely that such a process could be made to work in the case of a black hole, for the magnetic dipole moment of a collapsing body should be effectively swallowed in a very short space of time by the resulting black hole, the external dipole field being radiated away (ref. 2 and unpublished work of W. Israel). The coupling we shall now describe between the angular momentum of the black hole and that of the external world depends on a somewhat contrived process, but once it has been established that such a process is theoretically possible, it becomes reasonable to ask whether processes of this general kind might occasionally even occur naturally. (An even more contrived process is described in ref. 3.)

The process we suggest depends on a property of Kerr's solution of Einstein's vacuum equations[4]. This solution describes a rotating black hole, and is interesting because there are strong indications that it may in fact represent the general limiting situation of a black hole, any additional asymmetries being rapidly eliminated by gravitational radiation (ref. 2 and unpublished work of B. Carter). Our process depends on the existence of a region between the two surfaces known as the stationary limit and the (absolute) event horizon. The stationary limit is a surface at which a particle would have to travel with the local light velocity in order to appear stationary to an observer at infinity; and just inside which no particle (its velocity having to be bounded by the local light velocity) can remain stationary as viewed from infinity. The event horizon is the effective boundary of the black hole inside which no information can escape to the outside world. The stationary limit lies outside the event horizon in the Kerr solution touching it only where they intersect the rotation axis. (The stationary limit and the event horizon coincide in the case of the Schwarzschild solution, which is the limiting case of the Kerr solution for zero angular momentum).

To define the stationary limit more precisely, we consider the Killing vector (say, $\partial/\partial t$) which generates the time-translation symmetry of the Kerr solution. This becomes null at the stationary limit and spacelike within it. The Killing vector $\partial/\partial t$ is uniquely defined by the property that it is timelike with unit norm at infinity. The mass–energy of a test particle with four-momentum p_a is defined to be its scalar product with $\partial/\partial t$, namely the t-component, p_0. This quantity is conserved in collisions, is conserved along the world line of a freely moving particle between collisions, and agrees with the ordinary definition of mass–energy at infinity, so unquestionably it does describe the energy of a test particle correctly (as measured from infinity). The significance of the region between the stationary limit and the event horizon is that because $\partial/\partial t$ is spacelike there, it is possible for the energy p_0 to be negative in this region even though the vector p_a may be timelike (or null) and future pointing (as it must be for a real particle). Inside the event horizon this can still be true, but is of little value to us because the effects cannot be observed from outside.

Our process is now as follows. A particle (0) with four-momentum $\overset{(0)}{p}_a$ is dropped into the relevant region. It then splits into two particles 1 and 2, with

$$\overset{(0)}{p}_a=\overset{(1)}{p}_a+\overset{(2)}{p}_a$$

so that the mass–energy (as measured from infinity) of one of them is negative

$$\overset{(1)}{p}_0<0$$

and such that the other escapes back to infinity. The first of these is then "swallowed" by the black hole (it crosses the event horizon, in other words) while the second carries more mass–energy back to infinity than the original particle possessed

$$\overset{(2)}{p_0} > \overset{(0)}{p_0}$$

Thus some of the energy of the black hole has been extracted in the process.

It will be necessary to go into some detail in order to show that such a process actually can be realized with physical particle orbits. We demonstrate this with an explicit example. It is convenient to choose coordinates for the Kerr metric so that it takes the form

$$ds^2 = dt^2 - dr^2 - d\theta^2 + 2a \sin^2\theta \, dr \, d\varphi - (r^2 + a^2) \sin^2\theta \, d\varphi^2$$
$$- \frac{2mr}{\Sigma} (dr - a \sin^2\theta \, d\varphi + dt)^2$$

where $\Sigma = r^2 + a^2 \cos^2\theta$ (ref. 5). We choose units so that $c=1$; then $m = GM$ and $a = J/M$, where M is the gravitational mass, G is Newton's constant of gravitation and J is the angular momentum of the "hole", and thus a is the specific angular momentum of the hole. The coordinates have been chosen so that for $a > 0$, test particles with $\varphi > 0$ revolve in the same sense as the hole spins, and so that when $a = 0$ the metric reduces to the Eddington–Finkelstein form of the Schwarzschild solution[6,7], t being an advanced time parameter.

The stationary limit and the event horizon occur at

$$r = m + \sqrt{m^2 - a^2 \cos^2\theta}$$

and

$$r = m + \sqrt{m^2 - a^2} = r_+$$

respectively. For a particle travelling on an equatorial geodesic $\theta = \pi/2$, and p^2 and p_2 are always zero (numbering the coordinates $x^0 = t$, $x^1 = r$, $x^2 = \theta$, $x^3 = \varphi$). Because both $\partial/\partial t$ and $\partial/\partial\varphi$ are Killing vector fields, not only the particle's energy $E = p_0$ but also its angular momentum $\Lambda = -p_3$ is a conserved quantity. In the equatorial plane the stationary limit is at $r = 2m$ and a physical (that is, non-spacelike and future-directed) test particle has negative energy E if and only if $\Lambda < 0$ and $r - 2m < -\Delta\mu^2 r/\Lambda^2$, where μ is the rest mass of the particle and $\Delta = r^2 - 2mr + a^2$. The least energy such a particle can have is

$$\frac{2ma\Lambda + \sqrt{\Delta r\{r\Lambda^2 + \mu^2 (r^3 + a^2 r + 2ma^2)\}}}{r^3 + a^2 (r + 2m)} \qquad (1)$$

Taking $a=9m/10$, a specific angular momentum not very different from that of the Sun, then the event horizon occurs at $r_+=m(1+\sqrt{19}/10)<3m/2$. Let the particle (0) have unit rest mass and unit energy (so it drops from rest at infinity). Suppose that at $r=1.75m$, (0) splits into particles (1) and (2)

$$1=\overset{(1)}{E}+\overset{(2)}{E},\qquad \overset{(0)}{\Lambda}=\overset{(1)}{\Lambda}+\overset{(2)}{\Lambda}$$

and also

$$\overset{(0)}{p^1}=\overset{(1)}{p^1}+\overset{(2)}{p^1}\tag{2}$$

From formula (1), the least energy a physical test particle can have at $r=1.75m$, obtained by putting $\mu=0$, is approximately $0.0872\Lambda/m$. Therefore a convenient permitted value for $\overset{(1)}{E}$ is

$$\overset{(1)}{E}=\overset{(1)}{\Lambda}/20m\tag{3}$$

Writing $\overset{(2)}{\Lambda}=\alpha m\overset{(2)}{E}$ and solving equations (2) and (3) for $\overset{(1)}{E}$ and $\overset{(2)}{E}$ we get

$$\overset{(1)}{E}=\frac{\overset{(0)}{\Lambda}/m-\alpha}{20-\alpha},\qquad \overset{(2)}{E}=\frac{20-\overset{(0)}{\Lambda}/m}{20-\alpha}$$

In order for (0) to get inside $r=2m$, it is necessary that $\overset{(0)}{\Lambda}<481m/180$. Therefore the numerator is positive, and to recover the largest possible amount of energy we must take the largest possible value of α less than 20. A physical test particle at $r=1.75m$ which ultimately escapes to infinity must have Λ less than approximately $2.96Em$. Hence it is convenient to take $\alpha=2{,}9$. In order to be able to satisfy $\overset{(0)}{p^1}=\overset{(1)}{p^1}+\overset{(2)}{p^1}$ at $r=1.75m$, $\overset{(0)}{p^1}$ must be small. Thus we choose $\overset{(0)}{\Lambda}$ so that (0) has nearly reached an apsis at $r=1.75m$. The choice $\overset{(0)}{\Lambda}=2.63m$ will do. With these values for α and $\overset{(0)}{\Lambda}$, we have $\overset{(1)}{E}=-0.0158$ and $\overset{(2)}{E}=1.0158$. The apsidal equation

$$\mu^2r^3\dot{r}^2=(E^2-\mu^2)r^3+2m\mu^2r^2+(a^2(E^2-\mu^2)-\Lambda^2)r+2m(\Lambda-aE)^2$$

for (0) gives $\overset{(0)}{p^1} \doteqdot -0.0341$. Choosing $\mu=0$ the apsidal equations for (1) and (2) give $\overset{(1)}{p^1} \doteqdot -0.0411$, $\overset{(2)}{p^1} \doteqdot 0.007$ and $\overset{(2)}{\mu} \doteqdot 0.306$. Thus (2) is moving outwards with more energy than (0) possessed and escapes to infinity. The particle (1) has negative energy and cannot escape outside $r=2m$; it finally falls to within the event horizon. The angular momentum of (2) exceeds that of (0) by $-\overset{(1)}{\Lambda}$, that is, approximately $0.316m$.

We do not know what upper limit for $\overset{(2)}{E}$ is attainable for processes of this kind. But if the efficiency of the process is to be gauged by the ratio of energy gained to angular momentum (because, after all, it is the rotational energy of the hole that we are proposing to extract), then we are constrained by formula (1) which gives us the inequality

$$2mr_+ \overset{(1)}{E} - a\overset{(1)}{\Lambda} > 0 \tag{4}$$

This also expresses the fact that the scalar product of $\overset{(1)}{p^a}$ with the null Killing vector at the event horizon must be positive when (1) crosses the horizon. Thus the energy gain $(-\overset{(1)}{E})$ cannot be greater than $a/2mr_+$ times the angular momentum gain $(-\overset{(1)}{\Lambda})$.

It is interesting to note the relation of this inequality to another phenomenon, namely the apparent fact that the size of the hole (as measured by the surface area of its event horizon) increases even though its mass M can decrease. The surface area of $r=r_+$ at $t=\text{constant}$ (or, indeed, of any other cross-section of the tube $r=r_+$) is

$$A = 8\pi m r_+$$

setting $\mathrm{d}m = -\overset{(1)}{E}$ and $\mathrm{d}(ma) = -\overset{(1)}{\Lambda}$ we retrieve the inequality (4) in the form $\mathrm{d}A > 0$. In fact, from general considerations one may infer that there should be a natural tendency for the surface area of the event horizon of a black hole to increase with time whenever the situation is non-stationary. Thus the ideal of maximum efficiency would appear to be achieved whenever this surface area increase is as small as possible. The particular example described here is somewhat inefficient in this respect.

We thank R. P. Geroch for discussions.

R. PENROSE

Department of Mathematics,
Birkbeck College,
London

R. M. FLOYD

Willesden College of Technology and
Department of Mathematics,
Birkbeck College,
London

Received December 16, 1970.

[1] Bardeen, J. M., *Nature*, **226**, 65 (1970).
[2] Israel, W., *General Relativity and Gravitation*, **2**, No. 1 (Plenum, in the press).
[3] Penrose, R., *Rivista del Nuovo Cimento, Serie* 1, **1**, 252 (1969).
[4] Kerr, R. P., *Phys. Rev. Lett.*, **11**, 237 (1963).
[5] Boyer, R. H., and Lindquist, R. W., *J. Math. Phys.*, **8**, 265 (1967).
[6] Eddington, A. S., *Nature*, **113**, 192 (1924).
[7] Finkelstein, D., *Phys. Rev.*, **110**, 965 (1958).

48

SPINOR CLASSIFICATION OF ENERGY TENSORS

Gravitation: problems, prospects (dedicated to the memory of A. Z. Petrov) (Russian), pp. 203–15, 358. *Izdat. "Naukova Dumka", Kiev,* 1972.

A later account of this Ricci-tensor classification appeared in *Spinors and Space-Time* Volume 2, Sections 8.7, 8.8 (Penrose, R. and Rindler, W., Cambridge University Press 1986) and is presented with more geometric explanation and, I hope, more clarity.

376 Spinor classification of energy tensors. *Gravitation: problems, prospects (dedicated to the memory of A. Z. Petrov) (Russian)*, pp. 203–15, 358

1. GENERAL THEORY OF SPINOR CLASSIFICATION

In view of Professor Petrov's pioneering work on the algebraic classification of vacuum Riemann tensors [3] (or Weyl tensors) and of the additional insights which a spinor approach to this problem subsequently yielded, [1], [2], [4] it would seem appropriate to present, here, an extension of the spinor treatment having sufficient generality that it applies also to the classification of Ricci tensors.

The most straight-forward way to give an algebraic classification (with respect to the Lorentz group) of a symmetric two-index covariant tensor L_{ab} would appear to be to treat it as a matrix and use the methods of matrix algebra. This approach has been used by various authors, the earliest known to me being that of to be completed later [6], [8]. The matrix approach does have certain disadvantages, however, when it comes to considering the question of the specialization of algebraic types into one another. The situation is similar to one which arose already in the standard Petrov classification of the Weyl tensor [3]. For example, while type I_d is obviously a limiting case of non-degenerate type I, ($I \rightarrow I_d$), it is less obvious, in the matrix formulation, that the specialization can be achieved via a non-null type II as an intermediate stage: $I \rightarrow II \rightarrow I_d$. In terms of the spinor classification [1], both specialization schemes are equally obvious, being expressed, respectively, as [1111] → [22] and [1111] → [112] → [22]. When applied to a tensor such as L_{ab}, however, the spinor treatment yields a classification which is not nearly so transparent as it was in the case of the Weyl tensor. But it does not appear that this is really a drawback of the spinor technique as applied to symmetric two-index tensors. It is rather that the spinor formalism reveals an essential complication which is already present for such tensors, but which is rather surprisingly absent in the case of the more complicated-looking four index Weyl tensor.

---

[1] This article was written originally during the author's stay at the University of Texas in 1963/64. It was intended as a portion of a larger joint work with F. A. E. Pirani and R. K. Sachs on classification schemes for Ricci tensors. However, this more extended paper was abandoned, partly owing to overlap of some of the portions with work by other authors. I am grateful to Professor Pirani and Sachs for giving me permission to publish my portion of this paper separately.

The spinor classification technique which I shall present has the advantage that it may be applied to any tensor or spinor quantity with any number of indices. However it involves some ideas taken from algebraic geometry which may be a little unfamiliar. These extra complications arise when one considers the classification of a symmetric spinor possessing both primed and unprimed indices, rather than (as was the case with the Weyl spinor) a symmetric spinor having one type of index only.

A motivation for considering the question of specializations is that often a algebraic condition on a tensor would imply, not that it belongs to a certain class α, but it belongs either to α or to the class of limiting cases of α i. e. to the class $\bar{\alpha}$, where $\bar{\alpha}$ denotes the closure of the set α in an appropriately defined sense. For example, the condition that L_{ab} should possess a real null eigen-vector is one which would cut across most classification schemes as they stand. If the notion of specialization is incorporated, however, then such conditions are easily expressed.

The particular choice of classification (into disjoint classes $\alpha, \beta, \ldots$) that I shall make, will be governed primarily by the following requirements.

(a) Each class α shall form a connected manifold.

(b) The closure $\bar{\alpha}$ of a class α shall intersect each other class β either in β itself, or else in the empty set.

(c) Each class α shall be invariant under local Lorentz transformations.

The requirement of connectedness in condition (a) is important for the idea of a specialization scheme to make complete sense. For otherwise, members of different components of a class might specialize in different ways and could not specialize into each other. The fact that α should form a manifold is, perhaps, less essential, but we would like to think that the «singular points» of a class and appendages of lower dimensionality should be considered as belonging to separate subclasses. Similarly, condition (b) seems desirable, since if $\beta \neq \bar{\alpha} \cap \beta \neq \varnothing$ then $\bar{\alpha} \cap \beta$ should be considered as a separate subclass: «those β's which are limiting cases of α's». Condition (c) is of course essential for physical meaningfulness.

Consider the classification of some tensor or spinor quantity. Without loss of generality we may write all indices in a covariant spinor forms. Tensor indices are translated to spinor form [1, 4] by means of a quantity $\sigma^{a}_{AB'}$, which is a Hermitian (2×2) matrix [1] for each value of a, satisfying

---

[1] Tensor indices range over the values 1, 2, 3, 4, and spinor indices over 1, 2, or over 1′, 2′. (Thus tensor and spinor indices are here being used in the conventional way, rather than as «abstract labels» according to «Battelle conventions» [2]. This is only because the latter would have entailed some rewriting of the manuscript.) The signature of g_{ab} is $(+ - - -)$.

$$g_{ab}\sigma^{a}_{AP'}\sigma^{b}_{BQ'} = \epsilon_{AB}\epsilon_{P'Q'}, \tag{1.1}$$

where ϵ_{AB}, $\epsilon_{P'Q'}$ are skew-symmetric Levi-Civita symbols (components 0,1; —1,0) and which are used for lowering spinor indices according to the rule

$$\xi^{\cdots\cdot}_{\cdots A} = \xi^{B\cdots\cdot}_{\cdots}\,\epsilon_{BA},\quad \eta^{\cdot}_{\cdot P'\cdot\cdot} = \eta^{\cdot\cdot Q'\cdot}_{\cdots\cdot}\,\epsilon_{Q'P'}.$$

We shall consider the classification of a spinor $\psi_{\cdots\cdot}$ which is separately symmetric in all its unprimed and primed indices [1]:

$$\psi_{AB\ldots LP'Q'\ldots T'} = \psi_{(AB\ldots L)(P'\ldots T')} \tag{1.2}$$

(irreducible under the Lorentz group). For an unsymmetrical $\psi_{\cdots\cdots}$ the classification problem reduces to that of the simultaneous classification of several spinors of the form (1.2), since any spinor is linearly expressible in terms of its symmetrical part, symmetrical parts of its contractions and ϵ spinors, e. g.

$$\begin{aligned}\alpha_{ABCP'Q'} &= \alpha_{(ABC)(P'Q')} + \frac{1}{2}\alpha^{X'}_{(ABC)X'}\epsilon_{P'Q'} + \\ &+ \frac{2}{3}\alpha^{Y}_{AY(P'Q')}\epsilon_{BC} + \frac{1}{3}\alpha^{Y\ X'}_{AYX'}\epsilon_{BC}\epsilon_{P'Q'} + \\ &+ \frac{2}{3}\alpha^{Y}_{YB(P'Q')}\epsilon_{AC} + \frac{1}{3}\alpha^{YX'}_{YBX'}\epsilon_{AC}\epsilon_{P'Q'} \\ &+ \frac{2}{3}\alpha^{Y}_{YC(P'Q')}\epsilon_{AB} + \frac{1}{3}\alpha^{YX'}_{YCX'}\epsilon_{AB}\epsilon_{P'Q'}.\end{aligned}$$

For simplicity, however, the simultaneous classification of several symmetric spinors will not be considered here.

We shall be interested, in the first instance, in a classification of spinors (1.2) which is invariant under multiplication by a non-zero scalar:

$$\psi_{\cdots\cdot} \to \lambda\psi_{\cdots\cdot};\quad \lambda \neq 0, \tag{1.3}$$

then the classification of $\psi_{\cdots\cdot}$ can be accomplished by examining the zeros of the form

$$\psi_{A\ldots LP'\ldots T'}\,\xi^{A}\ldots\xi^{L}\eta^{P'}\ldots\eta^{T'}, \tag{1.4}$$

for all complex ξ^{A}, $\eta^{P'}$. The coefficients of the various powers of $\xi^1, \xi^2, \eta^{1'}, \eta^{2'}$ are the quantities (1.2). Each of ξ^{A}, $\eta^{P'}$ can be used as the projective coordinates on a complex projective line. The expression (1.4) is then a quantity whose vanishing is a function of a point of each line, i. e. of the topological product $\mathfrak{B}$ of the two lines. The expression $\xi^{A}\eta^{P'} = (\xi^1\eta^{1'}, \xi^1\eta^{2}, \xi^2\eta^{1'}, \xi^2\eta^{2'})$ can be used as projective coordinates for the projective manifold $\mathfrak{B}$, which represents this topological product. The coordinates $x^{AP'}$ which are of the form

$$x^{AP'} = \xi^{A}\eta^{P'} \tag{1.5}$$

---

[1] Round brackets denote symmetrization.

form the locus $\mathfrak{B}$:

$$|x^{AP'}| = 0 \tag{1.6}$$

in the projective three-space whose projective coordinates are ($x^{11'}$, $x^{12'}$, $x^{21'}$, $x^{22'}$). Thus, $\mathfrak{B}$ is the quadric surface

$$x^{11'}x^{22'} = x^{12'}x^{21'}. \tag{1.7}$$

Any complex projective quadric surface is generated by two systems of straight lines. They are respectively given by (1.5) when $\eta^{P'}$ is constant and ξ^A varies or when ξ^A is constant and $\eta^{P'}$ varies. Clearly any two generators belonging to opposite systems will intersect in a unique point whereas no two distinct generators belonging to the same system can intersect.

Note that the vector equivalent of $\xi^A\eta^{P'}$:

$$n^a = \sigma^a_{AP'}\xi^A\eta^{P'} \tag{1.8}$$

is the *general complex null vector*, since (1.7) is simply

$$x^{AP'}x_{AP'} \equiv x^a x_a = 0 \tag{1.9}$$

where $x^a = \sigma^a_{AP'}x^{AP'}$. Thus, any non-zero complex null vector can be represented up to proportionality as a point of the quadric $\mathfrak{B}$. The *real* null vectors (corresponding to (1.8) when $\eta^{P'} = \pm\bar{\xi}^{P'}$) give a certain cross-section of $\mathfrak{B}$: its *reality section* $\mathfrak{R}$. The reality section will be a real quadric which is effectively an ordinary 2-sphere. This is the sphere of real null *directions* at a point — the «celestial sphere» — essentially a space-like cross — section of the null cone. (Corresponding future and past null directions are considered as the same direction.) Thus $\mathfrak{B}$ is the «complexified celestial sphere».

Now the locus

$$\psi_{A\ldots LP'\ldots T'}\xi^A \ldots \xi^L\eta^{P'} \ldots \eta^{T'} = 0 \tag{1.10}$$

is some complex curve μ on $\mathfrak{B}$. Thus our classification problem reduces to the problem of classifying curves on a projective quadric surface. The methods of algebraic geometry can then be used to achieve the classification. In the first instance, the curve can be broadly classified according to its orders (p, q) where p and q are the (correctly counted) number of respective intersections that μ has with a general generator one or the other system on $\mathfrak{B}$. In fact, it is clear from (1.10) that p and q are the respective numbers of unprimed and primed indices of ψ.... . For, if $\eta^{P'}$ is constant, (1.4) become a homogeneous polynomial of degree p in (ξ^1, ξ^2) which therefore has p roots (correctly counted) and similarly if ξ^A is kept constant and $\eta^{P'}$ varies, we get q roots. Note that the (1,0) curves and (0,1) curves are the generators of the two systems on $\mathfrak{B}$. (The total projective order of μ is always the sum $p + q$.)

Secondly, the question of reducibility of the curve μ arises, i. e. is it the sum of two or more algebraic curves and, if so, what are their orders? Reducibility of μ corresponds to reducibility of the form (1.4). Clearly the curve μ can only degenerate into two curves of orders (r, s), (t, u) if $r + t = p$ and $s + u = q$. Note, in particular, that every $(p, 0)$ and $(0, q)$ curve is reducible to a set of generators of $\mathfrak{B}$. In the case (4,0) the complete reducibility of (1.4) to linear factors is the basis of spinor classification of the Weyl tensor.

Thirdly, the classification can be refined by examining the multiple points of the curve (nodes, cusps, etc.) if any. Fourthly, questions of reality are also of great importance here. That is, we shall be interested in knowing what is the nature (in general topological) of the intersection of μ with $\mathfrak{R}$, i. e., of the set of real points of μ.

2. CLASSIFICATION OF THE ENERGY TENSOR

The spinor equivalent of L_{ab} is

$$L_{ABC'D'} = \sigma^a_{AC'}\sigma^b_{BD'}L_{ab}. \tag{2.1}$$

The trace-free part M_{ab} of L_{ab},

$$M_{ab} = L_{ab} - \frac{1}{4} L^c_c g_{ab}, \tag{2.2}$$

has a symmetrical spinor equivalent:

$$M_{ABC'D'} = L_{(AB)(C'D')} = L_{ABC'D'} - \frac{1}{4} L^{xy'}_{xy'} \epsilon_{AB}\epsilon_{C'D'} =$$

$$= \sigma^a_{AC'}\sigma^b_{BD'}M_{ab} \tag{2.3}$$

and can therefore be classified according to the method of section 1.

Now, since $M_{ABC'D'}$ has two primed and two unprimed indices, the curve μ whose equation is

$$M \equiv M_{ABC'D'}\xi^A\xi^B\eta^{C'}\eta^{D'} = 0 \tag{2.4}$$

will be a (2,2) curve on $\mathfrak{B}$ (an elliptic quartic). When $\eta^{C'} = \bar{\xi}^{C'}$, we get the real points of the complex curve μ, namely $\mu \cap \mathfrak{R}$. The curve μ will have a real equation since M_{ab} is real ($M_{ABC'D'}$ is Hermitian). This equation is

$$M_{ab}x^a x^b = 0. \tag{2.5}$$

where x^a is null ($x^{AB'} = \xi^A\eta^{B'}$) and so satisfies

$$g_{ab}x^a x^b = 0. \tag{2.6}$$

By virtue of (2.6) we could replace (2.5) by the «more general» equation

$$L_{ab}x^a x^b = 0. \tag{2.7}$$

In fact μ is the general intersection of two projective quadrics (2.6) and (2.7) Equation (2.6) is simply the equation of $\mathfrak{B}$ — or with real coordinates, of the sphere $\mathfrak{R}$. For general L_{ab}, the curve μ is the general quartic curve with real equation on the sphere $\mathfrak{R}$. The real part of μ can consist of two separate pieces, of one piece, or of no piece at all. In more special cases μ can acquire a double point or degenerate into several curves of lower orders.

To obtain the equation of μ in an explicit form, put $x^1 = u$, $x^2 = v$, $x^3 = 1 - \frac{1}{4}(u^2 + v^2)$, $x^4 = 1 + \frac{1}{4}(u^2 + v^2)$. Then (2.6) is automatically satisfied and μ can be considered as having a Cartesian equation in the (u, v)-plane given by (2.5) or (2.7). In fact, u and v are conformal coordinates for the sphere $\mathfrak{R}$ which can be represented as $\left(1 + \frac{1}{4}(u^2 + v^2)\right)^{-1} \left\{u, v, 1 - \frac{1}{4}(u^2 + v^2)\right\}$.

Let us first consider the ways in which μ can degenerate. Being a (2,2) curve on $\mathfrak{B}$, we could envisage, in the first instance, a splitting up into (1,1) + (1,1), into (2,0) + (0,2), or into (2,1) + (1,0). The last possibility cannot arise here without a further splitting, however, since μ has a real equation. For the complex conjugate of a (2,1) curve is a (1,2) curve, so μ would also have to split up as (1,2) + (0,1). This would imply that the (2,1) part splits into (1,1) and (1,0) at least. If μ splits into (2,0) + (0,2), then the (2,0) part must be the complex conjugate of the (0,2) part. But here again, further splitting occurs, since every (2,0) curve splits (1,0) + (1,0) and similarly for (0,2). If μ splits (1,1) + (1,1) then two possibilities arise. Either each (1,1) curve has itself a real equation, or else each is the complex conjugate of the other. The further splitting of (1,1) into (0,1) + (1,0) is also possible. In every case the final result must be that all the irreducible parts either have real equations (and are $(p, p.)$) or they pair off into conjugate imaginary pairs ($(p, q) + \overline{(q, p)}$).

Thus, there are four possible types of component to consider, namely,

Q — irreducible (2,2) quadric with real equation,
C — irreducible (1,1) conic (circle) with real equation,
B — pair of complex conjugate irreducible conics $(1,1) + \overline{(1,1)}$,
X — pair of complex conjugate straight lines $(1,0) + \overline{(0,1)}$. The specialization possibilities are

$$\begin{aligned} &Q \to CC \\ &Q \to B \\ &B \to XX \\ &B \to C^2 \\ &C \to X. \end{aligned} \tag{2.8}$$

The distinction between CC and C^2 is that the symbol C^2 denotes a repeated conic while CC denotes a pair of distinct conics. Thus, if μ is C^2 we should have $M \equiv (F_{AC'}\xi^A\eta^{C'})^2$, i. e. $M_{AB}^{C'D'} = F_{(A}^{(C'}F_{B)}^{D')}$ for some Hermitian $F_{AC'}$, whereas if μ is CC, $M \equiv (F_{AC'}\xi^A\eta^{C'}) \times$

$\times (G_{BD'}\xi^B\eta^{D'})$ i. e. $M_{AB}^{C'D'} = F_{(A}^{(C'}G_{B)}^{D')}$ with $G_{BD'}$ Hermitian and distinct from $F_{BD'}$. Note that case B is given when $M \equiv (P_{AC'}\xi^A\eta^{C'}) \times \times (\bar{P}_{D'B}\xi^B\eta^{D'})$ i. e. $M_{AB}^{C'D'} = P_{(A}^{(C'}\bar{P}_{B)}^{D')}$, where $P_{AC'}$ is not Hermitian. Clearly C^2 is a limiting case of B. In case X, we have a form $(J_A\xi^A)(\bar{J}_{C}\eta^{C'})$. To get XX as a limiting case of B, put $P_{AC'} = J_AK_{C'}$. Then $M \equiv (J_A\bar{K}_C\xi^A\xi^C)(\bar{J}_{B'}K_{D'}\eta^{B'}\eta^{D'})$ i. e. $M_{ABC'D'} = = J_{(A}\bar{K}_{B)}\bar{J}_{(C'}K_{D')}$. (For general case B, $P_{AC'}$ should be restricted *not* to split in this way.) It may be remarked, also, that when μ is CX we get $M_{ABC'D'} = J_{(A}F_{B)(C'}\bar{J}_{D')}$ clearly a limiting case of CC. Of course, specialization

$$\begin{aligned} &CC \rightarrow C^2 \\ &XX \rightarrow X^2 \end{aligned} \qquad (2.9)$$

are also permitted as are any which could be compounded of specialization (2.8), (2.9), e. g.

$$Q \rightarrow CX,\ B \rightarrow X^2$$

(but clearly not $B \rightarrow CX$).

This classification scheme is as yet much too crude to satisfy the conditions laid down in section 1. Consider the case C^2, for example. The conic C can either have real points or not (according to whether it is a circle of real or imaginary radius). In order to pass continuously from the real case to the imaginary case, we must pass through the case X^2 (X is a «circle» of zero radius). Thus the manifold of type C^2 is not connected. A similar situation holds for type CX and CC.

In order to remedy the situation, it will be necessary first to make a sub-classification according to the number of distinct real pieces which are present for each component. Thus, type Q can be separated into three subtypes Q_2, Q_1, Q_0 and C separates into two subtypes C_1, C_0. The suffix in each case denotes the number of one-dimentional real parts possessed by the curve. (The nodal quartic (which resembles the symbol «∞» will be considered as a Q_1 curve, etc.) It will also be necessary to subdivide the types still further according to the real multiple point structure of μ. (Fortunately the complex multiple point structure need not be examined as it turns out to be determined by the reducibility and the real multiple point structure.) When this is done, the classification scheme obtained will satisfy all the conditions (a), (b), (c) of section 1. There are nine possible real multiple point structures to consider, namely [1]

n — real node with two real branches (double point);

i — isolated real node but with conjugate imaginary branches (double point);

nn — two real nodes with real branches;

ii — two isolated real nodes;

---

[1] See ref. [7], p. 57, for example, for a picture of Several each one of these multiple points (n: fig. 2.1; i: fig. 2.2; c: fig. 2.3; r: fig. 2.4; t: fig. 2.6).

c — cusp (double point with one branch and coincident tangents);
τ — tacnode (double point with two branches — real or imaginary — and coincident tangents);
t — triple point (with one real and two conjugate imaginary tangents);
τ^{∞} — real curve of «tacnodes» i. e. a repeated real curve;
q — quadruple point (with two repeated conjugate imaginary tangents).

The specialization scheme for the real multiple point structure is given in fig. 1, where the vacant space at the top denotes the absence of real multiple points.

The various possible types are now as follows.

Dimension

$$\begin{aligned}
&8 - Q_2, Q_1, Q;\\
&7 - Q_1i, Q_1n, Qi;\\
&6 - Q_1c, Bii, C_1C_1nn, C_1C_1, B, C_1C, CC;\\
&5 - B\tau, C_1C_1\tau, C_1Xi, CXi;\\
&4 - C_{1_2}Xt, XXii;\\
&3 - C_1\tau^{\infty}, C^2;\\
&2 - X^2q.
\end{aligned}\tag{2.10}$$

For conciseness and for later convenience, any suffix zero has been omitted from the symbols Q and C. Thus, Q and C above stand for what we had earlier denoted by Q_0 and C_0, respectively. The dimension figure at the left of (2.10) denotes the number of degrees of freedom of the curve μ for each specified type. M_{ab} has one more degree of freedom in each case (since it is determined by μ up to proportionality) and L_{ab} has just two more degrees of freedom. There is also the case of 0, not listed in (2.10), for which $M_{ab} = 0\,(L_{ab} \propto g_{ab})$. In this case there is no curve μ (or we could say $\mu = \mathfrak{B}$).

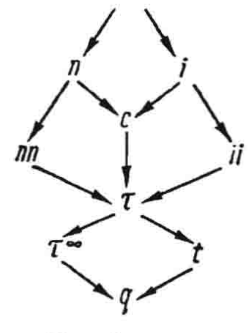


Fig. 1

A few words of explanation for the list (2.10) are perhaps called for. Any quartic curve on $\mathfrak{B}$ which possesses more than one double point, or a tacnode, or triple or quadruple point must necessarily be reducible. Hence, we have Bii rather than Qii, etc. Furthermore, when μ is reducible, there must be such a multiple point structure present, since the multiple points will be just the intersection points of the various components. However, these multiple points will not necessarily appear in the list (2.10) since in many cases they will occur in conjugate imaginary pairs (e. g. in the case CC there are two conjugate imaginary double points present). A tacnode occurs when two conics touch; a triple point occurs when a pair of conjugate imaginary lines intersect on a conic; an example of an isolated node is the isolated intersection point of a pair of conjugate imaginary lines («point circle»); A quadruple point occurs at the intersection point of two pairs of coincident lines (repeated «point circle»).

The condition for $x^{AB'} = \xi^A\eta^{B'}$ to be a multiple point of μ is

$$M_{ABC'D'}\xi^A\eta^{C'} = \lambda\xi_B\eta_{D'}, \tag{2.11}$$

for some λ (otherwise the left-hand side of (2.11) would determine the tangent to μ at x^a). For a real multiple point $\eta^{C'} = \bar{\xi}^{C'}$, whence the Hermiticity of $M_{ABC'D'}$ allows us to write (2.11) as

$$M_{ABC'D'}\xi^A\xi^B\bar{\xi}^{C'} = 0. \tag{2.12}$$

Note that (2.11) is the condition that the null vector x^a be an eigen-vector of M_{ab} — or equivalently, an eigen-vector of L_{ab}. Thus the real null eigen-vectors of L_{ab} correspond precisely to the real multiple points of μ.

The condition for a cusp (or tacnode or more special) is that (2.11) should hold ($\eta^{C'} = \bar{\xi}^{C'}$) and in addition,

$$2\lambda = |\nu|, \tag{2.13}$$

where

$$M_{ABC'D'}\xi^A\xi^B = \nu\bar{\xi}_{C'}\bar{\xi}_{D'}. \tag{2.14}$$

The existence of relation (2.14) follows from (2.12). The condition for an isolated node is that (2.13) be replaced by $2\lambda > |\nu|$ and for a node with real branches by $2\lambda < |\nu|$. For a triple point (or quadruple point), we have both of

$$M_{ABC'D'}\xi^A\xi^B = 0, \tag{2.15}$$

$$M_{ABC'D'}\xi^A\bar{\xi}^{C'} = 0 \tag{2.16}$$

holding. For a quadruple point,

$$M_{ABC'D'}\xi^A = 0. \tag{2.17}$$

Equation (2.15) by itself implies that μ splits into a pair of conjugate imaginary lines intersecting at $x^{AB'} = \xi^A\bar{\xi}^{B'}$ and another curve (i. e. $M_{ABC'D'} = \xi_{(A}F_{B)(C'}\bar{\xi}_{D')}$). Equation (2.16) by itself implies that $x^{AB'} = \xi^A\bar{\xi}^{B'}$ is a real node at which the two branches intersect orthogonally (regarding $\mathfrak{R}$ as a sphere) — or more special. It would have been possible to subdivide our types still further if we had considered such a node as being a distinct type of double point — this corresponds to the case when M_{ab} is *annihilated* by a null vector. However, an infinite variety of further subdivision is always possible, since in general, μ possesses two real projective invariants [1] (on $\mathfrak{B}$) one of which survives if μ acquires one or two nodes. These invariants could be specified as taking any value as desired. The invariants are essentially the ratios of the eigenvalues of M_{ab}.

---

[1] One of these, the *modulus* of the curve μ is also a birational (i. e. intrinsic) invariant of μ. It is essentially the cross-ratio of the four generators of $\mathfrak{P}$ of one or the other system, which *touch* μ. This cross-ratio is also the same as the cross-ratio associated with the principal null directions of the spinor $M^{P'Q'}_{(AB}M_{CD)P'Q'}$. (This spinor has some interest in that Plebanski [5] has used its classification, in terms of its principal null directions, to obtain a classification scheme for $M_{ABP'Q'}$.)

We now come to the complete picture of specialization for the list (2.10). This is given in fig. 2. The diagram is somewhat complicated and it is fortunate that the whole specialization scheme can, in fact, be summarized in a few comparatively simple rules.

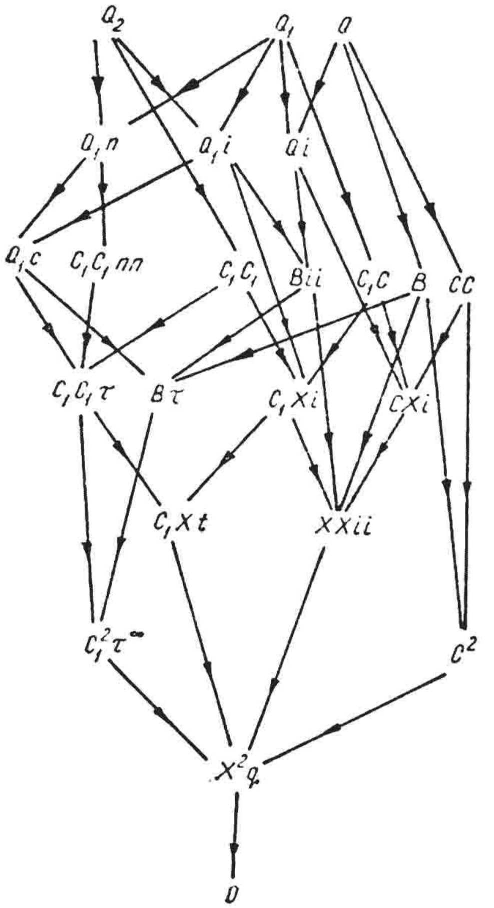


Fig. 2

(i) A specialization is allowed only if it is permitted by the general scheme (2.8) and (2.9) for the components and by fig. 1 for its real multiple point structure, and it reduces the dimensionality.

(ii) If a component possesses no one-dimensional real part, then the corresponding specialized components can possess no one-dimensional real part which is not double. (Effectively this means that no numerical suffix can appear in the specialization of a component if none is present already— with the exception that $C_1^2\tau^\infty$ is a limiting case of $B\tau$.)

(iii) The total number of one-dimensional real parts (i.e. the sum of the suffix numbers) cannot change in a specialization unless the multiple point structure becomes more special. Furthermore, it cannot change by more than unity (either way) in a specialization (except if the specialization is achieved in more than one step). For the purposes of this rule, $C_1^2\tau^\infty$ is considered as having a single real part.

The specializations allowed by fig. 2 (and those compounded of such specializations) are then just those not forbidden by rules (i), (ii), and (iii), together with the final specialization $X^2q \to 0$.

It is perhaps worth mentioning the modification of fig. 2 that is necessary if we are interested in an absolute classification of M_{ab} and not just a classification of the equivalence classes of M_{ab} up to proportionality as has been done up to now. Effictively this means a classification of M_{ab} which takes into account its sign. In most cases it is not possible to change M_{ab} continuously into $-M_{ab}$ without specializing its type at some stage. The sign of M_{ab} can be interpreted in terms of two regions into which the sphere $\mathfrak{R}$ is divided by the real points of the curve μ. One region will be called the positive region — corresponding to the real null vectors x^a for which $M_{ab}x^ax^b > 0$ and the other region, the negative region — corresponding to $M_{ab}x^ax^b < 0$. The sign of the region is reversed whenever the curve μ is crossed — except that the sign does not change where μ is double. (Thus, in case $C_1^2\tau^\infty$, the sign is the same on both sides of μ.)

Now, for each type, we define a positive and a negative version as follows.

1). In the cases Q, Qi, Bii, ... and also $C_1^2\tau^\infty$ where no real non-multiple part of μ is present, the positive versions $Q+$, $Qi+$, $Bii+$, ..., $C_1^2\tau^\infty+$ are those in which the whole sphere $\mathfrak{R}$, except for the multiple points of μ, is a positive region. Similarly, for the negative versions $Q-$, $Qi-$, ..., essentially the whole of $\mathfrak{R}$ is a negative region.

2). In cases Q_2, C_1C_1, Q_1i, C_1Xi one of the two regions is doubly connected (perhaps only because a point is missing), but not the other (which in the first two cases is disconnected). Q_2^+, ..., C_1xit+ have positive doubly connected regions, while Q_2^-, ..., C_1xi- have negative doubly — connected regions.

3). One of the two regions for cases Q_1n and $C_1C_1\tau$ is disconnected. This disconnected region is positive for Q_1n- and $C_1C_1\tau+$ and negative for Q_1n+ and $C_1C_1\tau-$.

4). Of the two regions for case Q_1C, one is «thorn-like» and possesses a «sharp point»; the other is «heart-like». The «heart-like» region is positive for Q_1C+ and negative for Q_1C-.

5). In cases Q_1, C_1C, C_1C_1nn, C_1Xt, 0 the positive versions cannot be distinguished from the negative versions.

The complete specialization diagram can now be constructed as follows. Take two copies of fig. 2, a positively labeled copy and a negatively labeled copy. Identify Q_1+, C_1C+, C_1C_1nn+, C_1xt+, $O+$ respectively with Q_1-, C_1C-, etc. in accordance with (5) above. Replace the incorrect specializations $C_1C_1+ \rightarrow C_1C_1\tau+$, $C_1C_1- \rightarrow C_1C_1\tau-$ by the correct specializations $C_1C_1+ \rightarrow C_1C_1\tau-$, $C_1C_1- \rightarrow C_1C_1\tau+$. The entire specialization scheme is now complete.

3. SOME PHYSICAL IDENTIFICATIONS

Suppose L_{ab} is to be an energy-momentum tensor. Then on physical grounds we require

$$L_{ab}t^at^b \geqslant 0 \tag{3.1}$$

for all time-like t^a. We may further impose a condition such as

$$L_a^a \geqslant 0 \tag{3.2}$$

on the trace of L_{ab}. Clearly (3.1) implies

$$L_{ab} n^a n^b \geqslant 0 \tag{3.3}$$

for all null n^a, i. e.

$$M_{ab} n^a n^b \geqslant 0 \tag{3.4}$$

for all null n^a. This in turn implies

$$M_{ab} t^a t^b \geqslant 0 \tag{3.5}$$

for all time-like t^a, since (3.5) can be obtained by adding six versions of (3.4) and using $M_{ab} g^{ab} = 0$, in which n^a is, in turn, $t^a \pm x^a$, $t^a \pm y^a$, $t^a \pm z^a$, where x^a, y^a, z^a, t^a form a pseudo-orthonormal tetrad. Clearly (3.5) and (3.2) together imply (3.1). Thus, we can take (3.4) and (3.2) to be physically reasonable conditions to be satisfied.

Now (3.4) is easily interpreted in terms of our refined classification scheme. The types which satisfy (3.4) are

$$Q+, Qi+, CC+, B+, Bii+, CXi+, B\tau+, XXii+, C^2+,$$
$$C_1^2\tau^\infty+, X^2q+, 0+ \tag{3.6}$$

namely the positive versions of those types possessing no numerical index, together with $C_1^2\tau^\infty+$ (i. e. every real point of μ must be multiple). The types which satisfy (3.4) with strict equality (with $n^a \neq 0$) are

$$Q+, CC+, B+, C^2+ \tag{3.7}$$

in which cases μ has no real point at all. In all cases in (3.6) except for type 0 with $L_a^a = 0$, the strict equality also holds in (3.1) (where $t^a t_a \geqslant 0$).

The cases of L_{ab} for dust or for perfect fluid is of some interest. In these cases the type is C^2+ since $L_{ab} = T_a T_b + \alpha g_{ab}$ whence $M_{AB}^{C'D'} = T_{(A}^{C'} T_{B)}^{D'}$, t_a being time-like. An electromagnetic field can be of type $XXii+$ or X^2q+ according as it is non-null or null since $M_{ABC'D'} = \Phi_{AB}\overline{\Phi}_{C'D'}$, where $\Phi_{AB} = J_{(A}K_{B)}$ or $\Phi_{AB} = J_A J_B$ respectively. The multiple points correspond to the principle null directions of the field.

Another interesting question is the asymptotic structure implied for the energy-momentum tensor of radiation (either electromagnetic or neutrino) when it is assumed that the appropriate «peeling off» property holds for the field. In this case we get, making appropriate assumptions,

$$L_{ab} = M_{ab} = v^{-2} M_{ab}^{(1)} + v^{-3} M_{ab}^{(2)} + v^{-4} M_{ab}^{(3)} +$$
$$+ v^{-5} M_{ab}^{(4)} + 0\,(v^{-6}),$$

where each $M^{(\alpha)}_{ab}$ is constant along a null ray with an affine parameter r. We have, from the «peeling off» property,

$$M^{(4)}_{ABC'D'}\xi^A\xi^B\bar{\xi}^{C'}\bar{\xi}^{D'} = 0,$$

$$M^{(3)}_{ABC'D'}\xi^A\xi^B\bar{\xi}^{C'} = 0,$$

$$M^{(2)}_{ABC'D'}\xi^A\xi^B = 0, \; M^{(2)}_{ABC'D'}\xi^A\bar{\xi}^{C'} = 0,$$

$$M^{(1)}_{ABC'D'}\xi^A = 0,$$

where ξ^A corresponds to the null direction of the null ray. We see from (2.4), (2.12), (2.15), (2.16), (2.17) that these are the respective conditions on $M^{(\alpha)}_{ABC'D'}$ ($\alpha = 1, 2, 3, 4$) for appoint of multiplicity at least $5 - \alpha$ corresponding to the direction of the null ray.

REFERENCES

1. Penrose R.— Annals of Physics, 1960, **10**, 171.
2. Penrose R. Structure of Space — Time, in Battelle Rencontres, ed. Witt C. M. de and Wheeler J. A. W. A. Benjamin Inc., New York, 1968.
3. Petrov A. Z.— Sci. Not. Kazan State University, 1954, **114**, 55.
4. Pirani F. A. E. Introduction to Gravitational Radiation Theory in Lectures in General Relativity Theory, Brandeis Summer Institute, Prentice—Hall, Inc,. Englewood Cliffs., 1964.
5. Plebanski J.— Acta Physica Polonica, 1964, 26, 963.
6. Churchill R. V.— Trans. Amer. Math. Soc., 1932, 34, 784.
7. Walker R. J. Algebraic Curves. Princeton University Press, Princeton N. J., 1950.
8. Petrov A. Z. The Type of the Field and the Type of the Energy Momentum Tensor. Perspectives in Geometry and Relativity, Ed. B. Hoffman; U. B. Bloomington, 1966.

49

BLACK HOLES AND GRAVITATIONAL THEORY

Nature **236(5347)** (1972), 377–80

Black Holes and Gravitational Theory

ROGER PENROSE

Department of Mathematics, Birkbeck College, London

There have been several recent claims to have detected black holes. In this article Professor Penrose describes the theoretical basis of black holes and some of the recent thinking on the subject.

AFTER about 10^{10} yr, according to accepted theory, our Sun will first expand and become a red giant (swallowing the Earth in the process) and then contract to a white dwarf—a star of mere planetary dimensions. Further contraction will be prevented by electron degeneracy pressure (a consequence of the exclusion principle as applied to electrons). As Chandrasekhar[1] showed in 1931, however, a non-rotating star of mass greater than about 1.3 $M_\odot$ (where $M_\odot$ is the mass of the Sun) cannot ultimately sustain itself in this way against further gravitational collapse. Stars of up to about 50 $M_\odot$ are known to exist, so the question of their eventual fate is a very real one. Although this question was largely answered by Oppenheimer and Snyder[2] in 1939, most astronomers paid little attention, apparently believing that in the final stage of collapse sufficient material would always be ejected to bring the mass of the resulting body down to below the Chandrasekhar limit—or to below the Oppenheimer–Volkoff[3] limit which is the corresponding maximum mass (variously denominated in the range 0.7 $M_\odot$–3 $M_\odot$) of a neutron star (far denser even than a white dwarf, the density being comparable with that inside an atomic nucleus) in which it is neutron degeneracy pressure which sustains the star against gravitational collapse. (Neutron stars are observed in pulsars, and for a mass of about 1 $M_\odot$, the radius of a neutron star should be roughly 10 km.) The Oppenheimer–Snyder picture was not, until recently, generally accepted because of its very strangeness. It is this picture which has come to be referred to as that of a black hole and which I shall be concerned with here.

No Escape

The characteristic feature of a black hole is that it represents a region of space into which the body has fallen, but from which no form of matter, energy, or information can escape. (A 1 $M_\odot$ black hole has a radius of about 3 km—only a little less than that of a neutron star.) The body itself will have fallen deep inside the black hole and possibly been crushed out of existence. Nevertheless, the black hole itself still exerts a gravitational pull on other bodies, so it is not totally unobservable. But the detection of black holes is difficult. It is only within the past year or so that claims have been seriously put forward that the invisible companions of certain stars belonging to binary systems (or of larger systems such as globular clusters) may indeed be black holes[4–15]. The detection of a black hole may be through its gravitational effect on the companion star, through its gravitational effect on material in the neighbourhood of one or both "stars", or through the possible emission of X-rays (or even of gravitational waves), as the surrounding material is inexorably, but very slowly, funnelled into the black hole. Alternative ("conventional") explanations seem still to be available for all these black hole candidates so far put forward, but this situation may not persist for long.

It should be pointed out, however, that the evidence which might seem most conclusive for the existence of a black hole, namely that there be a completely dark object of mass too great to be either a white dwarf or neutron star, is of rather negative character. Theory predicts that such dark "objects" (that is, black holes) should exist, but observations of this kind could hardly be regarded as very strong evidence for the theory. One could conceive that theory might really be very wrong, insofar as it forbids the existence of objects of the nature of white dwarfs or neutron stars that are too massive. If, however, an intensive search over several years fails to confirm the existence of dark "objects" of mass greater than 3 $M_\odot$, then this would be exceedingly uncomfortable for present theory.

I shall consider the black-hole concept and explore the theoretical reasons for believing that black holes should exist. Although most of the discussion of black holes is carried on within the framework of Einstein's general theory of relativity, the existence of a phenomenon of this type is not so very critically dependent on the particular theory of gravity which is used. Indeed, even within the framework of Newtonian theory it was predicted, as long ago as 1798 by LaPlace[16], that (as quoted by Eddington[17]) "the largest luminous bodies in the universe may, through this cause, be invisible to us". LaPlace's argument was that for sufficiently massive and condensed bodies the escape velocity at the surface would exceed the velocity of light. Thus all emitted light would be dragged back, rendering the object invisible. In particular, LaPlace points out that "A luminous star, of the same density as the Earth, and whose diameter should be 250 times that of the Sun, would not, in consequence of its attraction, allow any of its rays to arrive at us". It is worth noting that the formula expressing the condition for this to take place is $2mGc^{-2}>r$, in Newton's theory, where m and r are the mass and radius of the body, G the gravitational constant and c the speed of light. This is the same formula as that given by Einstein's theory, expressing the condition that the body should be trapped inside a black hole.

Not all theories of gravitation, however, predict the existence of such a phenomenon. In particular, the scalar gravitational theory of Nördstrom does not lead to a picture of this kind. This is because in Nördstrom's theory, light is not deflected in a gravitational field. In general, the black-hole phenomenon is a consequence of the fact that light is attracted by gravitating bodies—an observed effect (which, incidentally, can be regarded as observational disproof of Nördstrom's theory). The general principle involved—that in a powerful enough (spherically symmetrical) gravitational field, light cannot escape and inevitably falls in towards the centre of attraction—is largely independent of the details of the theory of gravitation used, provided that light is sufficiently attracted by gravity in a way similar to matter. The point is roughly that once the escape velocity exceeds that of light, then (assuming that special relativity holds locally so that the light velocity is the limiting velocity) nothing can escape. Indeed, in the theory of gravity which at the present time may be regarded as the most serious rival to general relativity, namely the (Jordan)–Brans–Dicke theory[18], black holes (essentially identical to those predicted by general relativity) develop[19,20].

I shall, for the moment, confine attention to general relativity and describe first the standard (Oppenheimer–Snyder) picture of spherically symmetrical gravitational collapse to a black hole. Then I shall consider the effect that deviations from spherical symmetry may be expected to have in the light of some recent theoretical results in general relativistic black-hole physics. Finally, I shall discuss some more conjectural aspects, in particular in relation to Weber's remarkable observations[21], apparently of gravitational waves emanating from the centre of our Galaxy.

To obtain the picture of spherically symmetrical gravitational collapse, we envisage matter falling radially inwards towards some central point and suppose that the pressures are not large enough to counter the gravitational pull. (Indeed, in general relativity, pressure itself is a source of extra gravitational attraction. When the matter is sufficiently concentrated, the additional gravitational pull due to pressure more than overcomes the direct effect of the pressure itself. Greater pressure then increases the tendency to collapse.)

After the matter has fallen to within a radius $r=r_m=2mGc^{-2}$ (the Schwarzschild radius) curious things happen, for light emitted from the surface of the matter falls inwards (Fig. 1). (At $r=r_m$, emitted light can "mark time", hovering for ever at the same distance from the centre.)

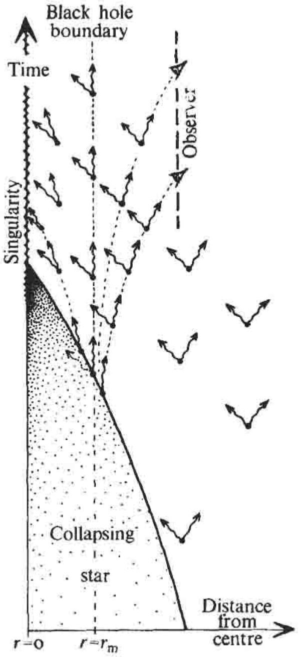


Fig. 1 The history of a spherically symmetrical star collapsing to form a black hole. The wavy arrows at each point indicate how an ingoing or outgoing light flash propagates.

It is clear from Fig. 1 that an observer outside the hole cannot see the matter after it has fallen in. In principle, however, he can always still see the matter just before it fell through the hole. This is because as the surface of the body approaches $r=r_m$, the light it emits takes longer and longer to escape, the escape time being theoretically infinite from $r=r_m$. But in practice the observer would receive essentially no light from the matter after a short time. This is because only a limited amount of light is emitted before the body reaches $r=r_m$, whereas this limited amount has to be received over an infinite time-span by the observer. Thus the observer sees the light increasingly redshifted, the intensity dying off very rapidly—exponentially, in fact, with a time-scale of the order of milliseconds for the case of a body of a solar mass or so. A black hole is then all that the observer perceives[22].

But what happens to the matter itself after it has fallen through the hole? According to the spherically symmetric general-relativistic picture, the matter inevitably gets crushed to greater and greater densities—without any limit. Within a finite time (milliseconds again, for a solar mass) the density becomes infinite, the curvature of space–time also becoming infinite so that a singularity in the space–time at the centre is the result[2].

How does such a space–time singularity manifest itself in physical terms? This can be answered by imagining an observer falling inwards through the hole, following the matter towards the centre. The effect of space–time curvature on the observer is that he experiences tidal forces. He feels himself to be stretched in one direction (namely in a line extending through himself to the central point) and to be squashed in the perpendicular directions. As he falls towards the centre this tidal effect increases and, indeed, for a solar mass this tidal effect would be easily enough to kill a man even before he reached the hole. But for larger masses (for example the 10^6 to $10^8\ M_\odot$ range that some astrophysicists believe would be appropriate for a black hole situated at the centre of the Galaxy), these tidal forces would be quite small until the observer is well within the hole. But as he nears the centre, the effect mounts inexorably. Not only is the observer himself soon destroyed by the mounting space–time curvature, but the very atoms of his body, even their constituent elementary particles, will themselves be ultimately squeezed out of existence.

But this picture assumes exact spherical symmetry. Because everything falls inwards to the same central point, the infinite densities and curvatures which result at that point may be argued to be atypical of a general collapse situation. Any asymmetries present in the initial state of the body would be expected to become enormously magnified as the body shrinks. Perhaps the different portions of the collapsing body would "miss" each other—even re-expand out of the black hole. For a time, it seems that hopes of this kind were entertained by various theoreticians[23,24]. More recent work has, however, shown that within the framework of general relativity (or the Brans–Dicke theory for that matter), such hopes cannot be realized. Rigorous theorems[25–27] imply that, once the matter has fallen to within a black hole, a space–time singularity of some form is inevitable—although its detailed nature may be very different from that arising in the spherically symmetrical case. (In the physical world, as opposed to mathematical models, we may take the phrase "space–time singularity" to mean a region in which space and time have become so locally distorted that the present laws of physics are no longer applicable.)

Whether or not any of the matter can conceivably ever re-emerge from a black hole region is somewhat less certain. But it could only do so if another even stranger phenomenon, a "naked singularity", occurs also. A naked space–time singularity is one which is visible to an external observer[28,29], unlike the one considered above (from which no signal could escape the black hole). The existence of naked singularities would present much more serious theoretical problems than singularities clothed by black holes. This is because physics at a space–time singularity is quite unknown. Extraordinary local effects (for example the production of matter out of gravitation) may be expected to occur. From the observational point of view these effects can be ignored if (and only if) information from the singularities cannot escape.

It is natural to ask, in view of such strange consequences, whether it is reasonable to suppose that matter can ever be sufficiently concentrated that a black hole develops. Might it not be that for some reason a collapsing star or superstar must always halt, or possibly reverse its collapse, before the critical size $r=r_m$ is reached? I think it is clear that although it is likely that in some cases the collapse can be halted (if there is sufficient rotation, for example), there will also be many situations in which a point of no return is passed[28]. In general, the larger the object or collection of objects involved, the less uncertainty there is that this point can be reached. For example, for a collection of bodies of the mass of a galaxy, it occurs at an average density less than that of air. Furthermore, this condition that a point of no return has been passed (existence of a "trapped surface"[25,27]) is one which is stable under perturbations of the initial conditions. Thus, provided the deviations from spherical symmetry are not too great, and the collapse has proceeded sufficiently far, a black hole (or naked singularity) must result.

Rotating Black Holes

If we assume that naked singularities are absent, then a rather precise picture of the final state of a black hole may reasonably be inferred[28-30]. This picture is based on Kerr's solution[31-33] of the Einstein vacuum equations describing a rotating black hole. The solution depends on just two parameters m and a, where m is the mass and $maGc^{-2}$ the angular momentum of the black hole. (If an electromagnetic field is also present, one further parameter e, describing the charge, occurs[34].) The picture that emerges is one of an asymmetrical, possibly rotating body collapsing down past the point of no return, and producing a black hole which at first possesses asymmetries resulting from those of the body. But in a short time these asymmetries are radiated away by gravitational radiation, only the parameters a, m remaining to describe the nature of the hole. Higher moments, such as quadrupole and octupole, all then take values explicitly dependent on a and m. (This picture is supported by perturbation analysis[35-39]. More persuasive, however, are results within the exact Einstein theory which almost establish that a black hole which has reached a stationary state must be a Kerr solution. Israel[40,41] showed this in the non-rotating (static) case, but required some extra conditions now effectively removed by unpublished work of H. Müller zum Hagen, D. Robinson and H. Seifert. Carter[42] effectively obtained the result for a rotating axi-symmetric black hole, although another family of solutions disconnected from that of Kerr is not yet ruled out. Hawking has shown that any stationary black hole must be either non-rotating or axi-symmetric, thus reducing the general case to those already treated. His original argument[30] was based on rather too strong an assumption concerning the nature of a stationary black hole, but an improved treatment[43] seems to have removed this difficulty.)

Gravitational radiation is the gravitational analogue of electromagnetic radiation (light). It travels with the same speed c as does light and its existence is a theoretical consequence not only of Einstein's theory but of almost any other relativistic theory of gravitation, the chief differences being that some theories, for example the Brans–Dicke theory, allow scalar waves as well as Einstein's tensor waves.

Tensor gravitational waves have apparently been detected by Weber, but not scalar waves[44]. This may be taken as some evidence against the Brans–Dicke theory, but it all depends on the mechanism by which nature produces Weber's waves. Hawking has pointed out[20] that in a collision between two black holes in the Brans–Dicke theory, the gravitational waves emitted are entirely tensor waves, as they are in standard general relativity. This is because the whole scalar part of the gravitational field must be radiated away (or swallowed) during the formation of a black hole in the Brans–Dicke theory, the result being a black hole identical to that of general relativity[19,20] (that is, presumably a Kerr solution).

In view of astronomical observations that violent releases of energy are taking place in various parts of the universe (for example, quasars, explosions in galactic nuclei, radio galaxies, the source of Weber's gravitational waves) it is reasonable to ask whether the mass-energy stored in a black hole might not in some circumstances be released to the outside world. This is not to suggest that the actual matter swallowed by the hole need in some way be released, but simply that substantial portions of the energy content of its mass be recoverable. In fact there are several processes, feasible in varying degrees, which can be invoked for energy extraction. The simplest of these is to allow black holes to coalesce and, in the process, to emit energy in the form of gravitational waves. Misner has indicated how the process of congealing a large number of black holes to a single one could, in principle, allow (proportionally) almost the entire original mass-energy to be emitted as waves[28]. There are also methods whereby the rotational energy of a black hole may in principle be extracted. One of these[28,45,46] is to allow a particle to fall from infinity into a region termed the "ergosphere" which lies just outside the surface of a rotating black hole. When in the ergosphere the particle has to split into two in such a way that one portion escapes to infinity possessing more mass-energy than the original particle. The other is swallowed by the black hole. If repeated many times, this process would lead to the slowing down of the rotation of the black hole and to a substantial reduction in its mass. The energy of this mass loss is recoverable (very efficiently in principle) in the kinetic energy of the final particles. (I am not suggesting that this process is likely to be important in realistic astrophysical situations, merely that if energy extraction is possible in principle, then it is conceivable that nature has found some means whereby the same ends may be achieved in fact.)

One characteristic aspect of gravitation physics is that whereas one normally finds that the extremely small size of the gravitational constant leads to an almost total insignificance of gravitational effects when compared with other forces in nature, there are sufficiently extreme situations when just the reverse is the case and the conversion of mass to energy is achieved gravitationally with an efficiency of almost 100%. It is thus worth examining any process, no matter how contrived, if it yields some hope of explaining the almost miraculous apparent efficiency of energy production in some astronomical phenomena.

A useful theoretical principle for the discussion of energy extraction from black holes is that of surface area increase. This principle has been used in various contexts[30,45] and states that the total surface area of the black hole boundaries can never decrease (and can apparently remain constant only in a stationary situation). The boundary of a black hole (the "event horizon") is defined as a surface such that signals cannot escape from within it, but can escape from points outside it. (In the spherically symmetrical case this surface is $r=r_m$.) Because the surface area of a Kerr black hole is easily computed (and is $8\pi m\ \{m+(m^2-a^2)^{\frac{1}{2}}\ \}G^2c^{-4}$) this leads rapidly to several interesting inequalities. For example, there is Christodoulou's[47] concept of irreducible mass m_{ir} given by $16\pi m_{ir}^2=8\pi m\{m+(m^2-a^2)^{\frac{1}{2}}\}$ for a Kerr black hole, this being the minimum mass to which one could reduce the black hole by energy extraction processes. Hawking[30] has also used the area principle to great effect in the study of black hole collisions. If, for example, two non-rotating black holes each of mass m congeal to form a single black hole of mass M, then $M>m\sqrt{2}$. If the two come together without initially having significant relative velocity, then one can infer that an upper limit to the energy of the emitted radiation is $(2-\sqrt{2}\ mc^2$. It should be pointed out that black holes never bifurcate, they can only congeal.

Another way that black holes could be involved in energy production is in the extraction of the mass-energy content of material bodies, stars for example. The possibility of a large black hole (of 10^4 to 10^8 $M_\odot$, say) residing at the centre of the Galaxy has often been suggested. It may be supposed that such a black hole would be continually swallowing stars.

As each star is swallowed, a certain proportion of its mass-energy would be expected to be radiated as gravitational waves. In addition, the tidal force on the star might be large enough to produce considerable "side effects" (like electromagnetic radiation and bifurcation of the star). What would happen depends a great deal on the details of the situation. Various authors have tried to use these ideas in order to explain the extraordinary observations made by Weber, which apparently imply a large flux of gravitational radiation coming from the centre of our Galaxy. This flux is so large, in fact, that if the natural assumption is made that the radiation emanates isotropically from the galactic centre, then an absurdity seems to result, namely that the Galaxy is losing mass at the rate of 2,000 $M_\odot$ a year[21]—or perhaps even 10^5 or 10^6 $M_\odot$ per year (P. Kafka, private communication). (At the 2,000 $M_\odot$ a year rate, if continued, the entire Galaxy would disappear in less than 10^8 yr, a mere 1% of its present age.) Field, Rees and Sciama[48,49] have pointed out that observations of star orbits have ruled out the possibility that such a mass loss from our Galaxy can have been continuing for anything like 10^8 yr. They argue that a loss of 70 $M_\odot$ a year is much more in keeping with observations. In view of this discrepancy (and setting aside the possibility that the galactic centre has become active only recently —or that Weber's observations do not measure what they seem to) we are led to consider the possibility that there are gravitational waves coming from the galactic centre which are strongly beamed in the plane of the Galaxy. It may be pointed out that the Earth lies very close to the galactic plane, about 0.1% of its distance to the galactic centre[50], and in this way a mass loss of, say, as little as 10 $M_\odot$ a year could be consistent with Weber's observations, and so consistent also with the calculations of Field, Rees and Sciama.

But what conceivable mechanism could be responsible for such beaming? An interesting suggestion by Misner (private communication) involves an "extreme" Kerr black hole, that is one for which $a=m$. Bardeen[51] has suggested that such extreme black holes may occur through the accretion of matter at galactic centres, the stars being assumed to spiral gradually inwards in ever quickening (co-planar) circles. According to the Misner mechanism, these spiralling stars will beam gravitational radiation ahead of them in a small angle, in a way analogous to the electromagnetic synchrotron radiation which is produced by relativistic charged particles in electromagnetic fields. The argument is that this beaming would be sufficient not only to confine the radiation effectively to the galactic plane but also to produce the high frequency (1,660 Hz) short pulses that Weber claims to detect.

There are severe difficulties involved, to my mind, in making such a model work, but it does seem to represent a serious attempt to explain Weber's observations on the basis of accepted physical theory. As such, its consequences should be explored in full detail to ascertain the total plausibility of the idea. It may be that other suggestions involving black holes will emerge as alternative explanations of Weber's waves. On the other hand, it seems quite possible that all such explanations will eventually founder. If so, there remains the possibility that some rather different kind of gravitational catastrophe may be responsible for Weber's waves. It has long seemed to me that the presumption that the space–time singularities that arise in gravitational collapse must inevitably reside inside black holes is a product largely of wishful thinking. The possibility that naked singularities may sometimes arise must be considered seriously. Many solutions of Einstein's equations with naked singularities exist; one example is the Kerr solution with $a>m$. (This solution does not represent a black hole, being beyond the "extreme" case $a=m$.) The problem, however, is whether or not such naked singularities are likely to be self-sustaining or to occur in the collapse from a perfectly reasonable non-singular initial state. If, as might then be expected, matter production or gravitational wave production could take place at or near such a singularity, the astrophysical consequences could be considerable. It is worth remarking, in this connexion, that a large Kerr solution with $a>m$ situated at the centre of the Galaxy would have the property that signals from the neighbourhood of its (naked) singularity are necessarily beamed in one plane[33]—and this could be the galactic plane if the singularity were in some way also responsible for the very existence of the galactic plane. Of course suggestions of this sort would be hard to substantiate in the absence of a detailed theory of singularities. But if all black-hole explanations fail, it may be that we shall be driven to consider possibilities of this kind.

[1] Chandrasekhar, S., *Astrophys. J.*, **74**, 81 (1931).
[2] Oppenheimer, J. R., and Snyder, H., *Phys. Rev.*, **56**, 455 (1939).
[3] Oppenheimer, J. R., and Volkoff, G., *Phys. Rev.*, **55**, 374 (1939).
[4] Cameron, A., *Nature*, **229**, 173 (1971).
[5] Woolf, N. J., *Nature*, **230**, 39 (1971).
[6] Denarque, P., and Morris, S. L., *Nature*, **230**, 516 (1971).
[7] Cameron, A., *Nature*, **231**, 148 (1971).
[8] Novikov, I. D., and Zel'dovich, Ya. B., *International Conference on Gravitation and Relativity* (Copenhagen, GR6 Nordita, July 1971).
[9] Peebles, P. J. E., *International Conference on Gravitation and Relativity* (Copenhagen, GR6 Nordita, July 1971).
[10] Hawking, S. W., and Gibbons, G., *Nature*, **232**, 465 (1971).
[11] Devinney, E. J., *Nature*, **233**, 110 (1971).
[12] Batten, A. H., and Olowin, R. P., *Nature*, **234**, 341 (1971).
[13] Gott, J. R., III, *Nature*, **234**, 342 (1971).
[14] Wilson, R. E., *Nature*, **234**, 406 (1971).
[15] Gott, J. R., III, *Astrophys. J.* (in the press).
[16] LaPlace, P. S., *Allgemeine Geographische Ephemeriden, verfassed von Einer Gesellschaft Gelehrten*, 603, IV B ds I St. 1799 (Weimer, May 1798).
[17] Eddington, A. S., *The Internal Constitution of the Stars* (Dover, New York, 1959).
[18] Brans, C., and Dicke, R. H., *Phys. Rev.*, **124**, 925 (1961).
[19] Thorne, K. S., and Dykla, J. J., *Astrophys. J.*, **166**, 135 (1971).
[20] Hawking, S. W., *Commun. Math. Phys.*, **25**, 167 (1972).
[21] Weber, J., *Phys. Rev. Lett.*, **25**, 180 (1970).
[22] Podurets, M. A., *Astr. Zh.*, **41**, 1090 (1964); English translation, *Soviet. Astron.—AJ.*, **8**, 868 (1965).
[23] Lindquist, R. W., and Wheeler, J. A., *Rev. Mod. Phys.*, **29**, 432 (1957).
[24] Lifshitz, E. M., and Khalatnikov, I. M., *Adv. Phys.*, **12**, 185 (1963).
[25] Penrose, R., *Phys. Rev. Lett.*, **14**, 57 (1965).
[26] Hawking, S. W., *Proc. Roy. Soc.*, A, **300**, 187 (1967).
[27] Hawking, S. W., and Penrose, R., *Proc. Roy. Soc.*, A, **314** 529 (1970).
[28] Penrose, R., *Rivista Nuovo Cim.*, Serie I, **1**, Numero Speciale 252 (1969).
[29] Thorne, K. S., *Comments Astrophys. and Space Phys.* (G.B.), **11**, 191 (1970).
[30] Hawking, S. W., *Commun. Math. Phys.*, **25**, 152 (1972).
[31] Kerr, R. P., *Phys. Rev. Lett.*, **11**, 237 (1963).
[32] Boyer, R. H., and Lindquist, R. W., *J. Math. Phys.*, **8**, 265 (1967).
[33] Carter, B., *Phys. Rev.*, **174**, 1559 (1968).
[34] Newman, E. T., Couch, E., Chinnapared, K., Exton, A., Prakash, A., and Torrence, R., *J. Math. Phys.*, **6**, 918 (1965).
[35] Regge, T., and Wheeler, J. A., *Phys. Rev.*, **108**, 1063 (1957).
[36] Doroshkevich, A. G., Zel'dovich, Ya. B., and Novikov, I. D., *Zhur. Eksp. Teor. Fiz.*, **49**, 170 (1965); English translation, *Sov. Phys.—J.E.T.P.*, **22**, 122 (1966).
[37] de la Cruz, V., Chase, J. E., and Israel, W., *Phys. Rev. Lett.*, **24**, 423 (1970).
[38] Vishveshwara, C. V., *Phys. Rev. D*, 2870 (1970).
[39] Price, R., *Phys. Rev.* (in the press).
[40] Israel, W., *Phys. Rev.*, **164**, 1776 (1967).
[41] Israel, W., *Commun. Math. Phys.*, **8**, 245 (1968).
[42] Carter, B., *Phys. Rev. Lett.*, **8**, 331 (1971).
[43] Hawking, S. W., and Ellis, G. F. R., *The Large-scale Structure of Space–time* (Cambridge University Press, 1972).
[44] Weber, J., *Nuovo Cim.*, **4B**, 197 (1971).
[45] Penrose, R., and Floyd, R. M., *Nature Physical Science*, **229**, 177 (1971).
[46] Ruffini, R., and Wheeler, J. A., in *The Significance of Space Research for Fundamental Physics* (ESRO, Paris, 1971).
[47] Christodoulou, D., *Phys. Rev. Lett.*, **25**, 1596 (1970).
[48] Field, G. B., Rees, M. J., and Sciama, D. W., *Comm. Astrophys. Space Sci.*, **1**, 187 (1969).
[49] Sciama, D. W., *Nature*, **224**, 1263 (1969).
[50] Gum, C. S., Kerr, F. J., and Westerhout, G., *Mon. Not. Roy. Astron. Soc.*, **121**, 132 (1960).
[51] Bardeen, J. M., *Nature*, **226**, 64 (1970).

50

IDEAL POINTS IN SPACE-TIME

Proc. Roy. Soc. Lond. Ser. A **327(1571)** (1972), 545–67. With R. Geroch and E. H. Kronheimer

I have frequently found that this way of defining singular points and points at infinity in terms of the internal causal geometry of a space-time to be very useful (see, for example, Chapters 96, 104, and 270). These are the TIPs and TIFs, which relate to the work of Chapter 32. This procedure is also very similar to an earlier one due to H.-J. Seifert (1971) *General Relativity and Gravitation*, **1**, 247. It is probably best not to try to identify the past and future boundaries, as is attempted here, and the procedure seems to be most satisfactory when strong cosmic censorship holds (see Chapter 270), since then these two boundaries can be kept separate.

Ideal points in space–time

BY R. GEROCH†
Department of Physics, Syracuse University, Syracuse, New York
AND E. H. KRONHEIMER AND R. PENROSE‡
Department of Mathematics, Birkbeck College, London

(*Communicated by H. Bondi, F.R.S. – Received* 22 *September* 1971)

A prescription is given for attaching to a space–time M, subject only to a causality condition, a collection of additional 'ideal points'. Some of these represent 'points at infinity', others 'singular points'. In particular, for asymptotically simple space–times, the ideal points can be interpreted as the boundary at conformal infinity. The construction is based entirely on the causal structure of M, and so leads to the introduction of ideal points also in a broad class of causal spaces. It is shown that domains of dependence can be characterized in terms of ideal points, and this makes possible an extension of the domain-of-dependence concept to causal spaces. A suggestion is made for assigning a topology to M together with its ideal points. This specifies some singular-point structure for a wide range of possible space–times.

1. INTRODUCTION

A striking feature of many models of space–time, constructed according to the general theory of relativity, is that they possess curvature singularities. In the 'big bang' models, for example, the universe arises from an initial singularity, while in collapse situations the system evolves to a singular state. Recent work (Penrose 1965*a*, 1968; Hawking 1966*a*, *b*, 1967*a*; Hawking & Penrose 1970; cf. Geroch 1970*b*) has shown that the existence of some form of singularity is, in fact, a characteristic feature of large classes of solutions of Einstein's equations. Thus there is apparently a limit to the domain of applicability of the equations of general relativity – or even of normal ideas of space–time. In order to discuss the physics 'at a space–time singularity', one would, presumably, require a much deeper understanding of the interrelation between quantum theory, general relativity, and particle physics than is available at present. But if we think of singularities simply as points of a *boundary* of a smooth space–time manifold M, we can at least ask how such a boundary should be constructed, and what kind of boundary can arise from physically reasonable space–times. Since the singularities themselves cannot be regarded as actually belonging to the manifold,[1]§ we are led to consider methods of constructing additional *ideal points* which, when adjoined to M, result in a unified structure $\bar{M}$ incorporating 'singular' as well as 'non-singular' points.

There is another context in which the concept of ideal points has already proved valuable, namely in the study of asymptotic properties of sufficiently well-behaved

---

† Present address: Enrico Fermi Institute, University of Chicago, Chicago, Illinois. Supported by A.R.L. and A.F.O.S.R. Work done in part at Birkbeck College, London.

‡ Elected F.R.S. 16 March 1972.

§ The affixed numbers refer to notes at the end of the paper (p. 566).

(e.g. asymptotically flat or, more generally, asymptotically simple) space–times (Penrose 1964, 1965*b*, 1968). In this case the $\bar{M}$ involved has the structure of a conformal manifold with boundary. Radiation properties of M may be described, without recourse to limiting procedures, in terms of the behaviour of fields at *conformal infinity*: that is, at the boundary of $\bar{M}$. Many of the subtle features of gravitational radiation have comparatively straightforward interpretations in terms of the structure of $\bar{M}$. However, in the original definition of conformal infinity, the points of $\bar{M}$ were merely postulated rather than being constructed from M. As a consequence, the uniqueness of $\bar{M}$, given M, was not self-evident.

We shall present a precise definition of ideal points for space–times. Some of these ideal points can be interpreted as singular points of M, others as points at infinity. Our construction can be applied provided only that no two distinct points of M share the same past or the same future (i.e. that M is both *past-* and *future-distinguishing*), a condition only marginally stronger than the requirement that M should contain no self-intersecting null or timelike curves (Kronheimer & Penrose 1967). This construction of $\bar{M}$ is manifestly *conformally invariant*: the same $\bar{M}$ arises if the metric of M is altered by the introduction of an arbitrary conformal factor. It is because of this conformal invariance that the points at infinity and singular points are introduced simultaneously. It is in the *interpretation* – not in the *construction* – of the ideal points that the particular choice of conformal factor becomes relevant.

Various methods for constructing ideal points have been put forward. One of these (Hawking 1967*b*; Geroch 1968) involves assigning to each incomplete, inextendible geodesic an ideal end-point. A procedure is then introduced for identifying these end-points when the geodesics in question approach, in an appropriate sense, the same ideal point. Alternatively (Seifert 1968, 1971) one assigns the end-points to all inextendible causal curves and identifies those having corresponding pasts and futures, using techniques somewhat similar to the ones to be described in this paper. Yet another method (Schmidt 1971) involves introducing a positive-definite metric on the bundle of orthonormal tetrads of M. The completion of this bundle, as a metric space, is then factored by the action of the Lorentz group on the fibres to obtain the original manifold M with ideal points attached. Our method does not in general yield the same information as these, but has the decided advantage that it is comparatively simple: in fact, $\bar{M}$ may readily be found by inspection in many examples.

Our definition of ideal points appears, moreover, to have some merit from the physical point of view. Each ideal point is here directly associated with the region of space–time which it can influence, or be influenced by. This feature is of particular interest in the case of singularities, for it is to be expected that a region of near-infinite curvature would be a source of locally observable effects. It is of importance, therefore, to be able to determine those regions of space–time from which such effects could, in principle, be observed.

We construct the ideal points (§ 2) using only the causal (and not, for example, the differentiable) structure of M. Consequently, the ideal points can be defined when M is a causal space (Kronheimer & Penrose 1967) – or indeed merely a set with a partial ordering – of a certain type (*semi-full*). Moreover, the ideal points make possible a definition of the concept of the domain of dependence (cf., for example, Courant & Hilbert 1965) applicable not only to space–times but also to causal spaces. § 3 is devoted to a discussion of ideal points and their applications entirely within the framework of causal-space theory.

In § 4 we apply the results of the earlier sections to some particularly awkward (but simple) models for singular space–times. This leads to a suggestion for an ideal-point topology which gives a reasonable structure to $\bar{M}$ in all cases we have been able to examine.

2. IDEAL POINTS OF SPACE–TIMES

Let M be a space–time.[2] A fundamental notion is that of the (chronological) *past* of a point p, denoted by $I^-(p)$ (cf. Kronheimer & Penrose 1967), and defined as the collection of all past end-points of time like curves[3] in M with future end-point p. The (chronological) *future* $I^+(p)$ of p is similarly defined. These sets are open in M. For example, the past of the origin of Minkowski space is (in the usual coordinates) the solid cone $x^2+y^2+z^2 < t^2$, $t < 0$. To avoid pathological situations (such as that of the Gödel universe, in which the past of each point is the entire space), we shall as already mentioned restrict our consideration (even when the fact is not explicitly stated) to *past-* and *future-distinguishing* space–times: that is, to those in which $p \neq q$ implies both $I^-(p) \neq I^-(q)$ and $I^+(p) \neq I^+(q)$. Neither the past nor the future of a point can then include the point itself.

The past, $I^-[S]$, of a set S in M is defined as the union of the pasts of the points of that set: that is, as the collection of all points which may be reached from *some* point of S by a past-directed timelike curve. Similarly for the future, $I^+[S]$, of S. Note that the past and future of any set are each unchanged if the metric is altered by the introduction of an arbitrary conformal factor.

The past of any timelike curve in the space–time M will be called an 'IP'. (This is an abbreviation for 'indecomposable past-set', a term we shall explain shortly.) The IPs so defined are divided into two classes, consisting of those which are of the form $I^-(p)$ for some $p \in M$ (called *proper* IPs, or PIPs), and those which are not (called *terminal* IPs, or TIPs). IFs, PIFs, and TIFs are defined in a similar way by interchanging the roles of past and future. The boundaries of the TIPs are essentially the *event horizons* of Rindler (1956) (cf. also Penrose 1968), while the boundaries of the TIFs are essentially the *particle horizons*. We denote by $\hat{M}$ the collection of all IPs of M, and by $\check{M}$ the collection of all IFs.

We illustrate these notions by exhibiting the collection $\hat{M}$ when M is Minkowski space. For each point p of M, the set $I^-(p)$ is a PIP since it is identical with the past of any timelike curve with future end-point p. (Note that many different timelike curves may determine the same IP.) The past of the entire t-axis (in the usual coordinates) is, evidently, the whole of Minkowski space; and so this constitutes a TIP by our definition. But there are other TIPs which are not the whole space: for example, the past of the timelike hyperbola

$$t = \sinh\lambda, \quad x = \cosh\lambda, \quad y = z = 0, \quad -\infty < \lambda < \infty.$$

The TIP defined by this curve is that part of M lying to the past of the null hyperplane $x = t$ (cf. figure 1). In fact, every TIP in Minkowski space (with the exception

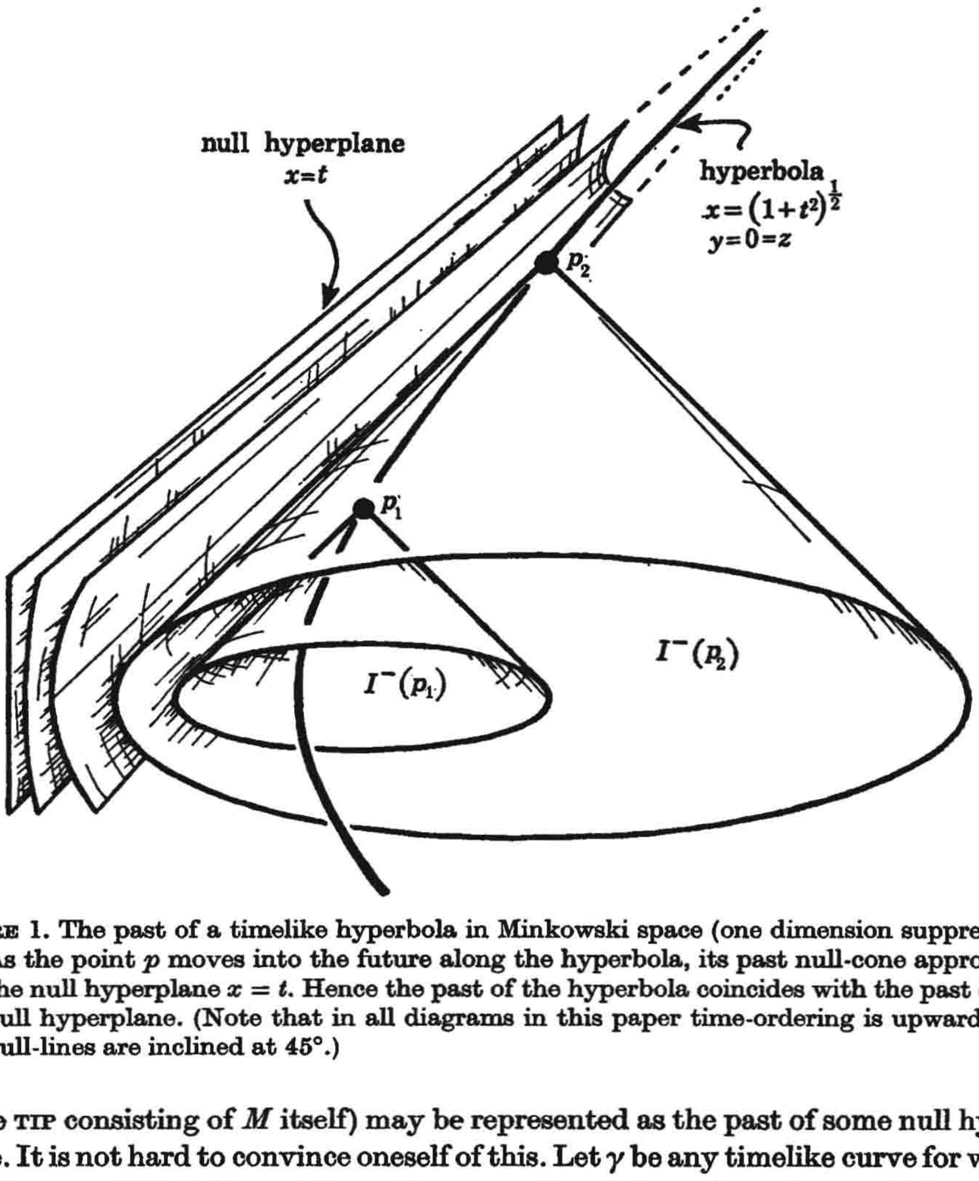


FIGURE 1. The past of a timelike hyperbola in Minkowski space (one dimension suppressed). As the point p moves into the future along the hyperbola, its past null-cone approaches the null hyperplane $x = t$. Hence the past of the hyperbola coincides with the past of the null hyperplane. (Note that in all diagrams in this paper time-ordering is upwards and null-lines are inclined at 45°.)

of the TIP consisting of M itself) may be represented as the past of some null hyperplane. It is not hard to convince oneself of this. Let γ be any timelike curve for which $I^-[\gamma]$ is a TIP. If $I^-[\gamma]$ is not the entire space, then it has a boundary, which may be regarded as the limit of the past light-cone $I^-(q)$ as q recedes to infinity in the future-direction along γ. It is clear that such a limiting past light-cone will be a null hyperplane. (For an alternative argument, see theorem (2.7).)

In figure 2, Minkowski space is drawn in such a way that its TIPs may be represented by points. Here $\mathbb{I}^+$ represents the TIP corresponding to the entire space–time, while the points of $\mathscr{I}^+$ represent the remaining TIPs. The TIPs themselves – i.e. the appropriate subsets of Minkowski space – are obtained by taking the pasts (in the obvious sense) of the points representing them in figure 2. Similar remarks apply, of course, to the TIFs. Note that, in this example, the TIPs and TIFs together essentially make up the boundary at conformal infinity (Penrose 1964, 1965*b*) which characterizes the asymptotic structure of Minkowski space. Thus our interpretation of $\check{M}$ [resp. $\hat{M}$] is that of Minkowski space, together with the points at past [resp. future] conformal infinity and the additional point $\mathbb{I}^-$ [resp. $\mathbb{I}^+$].

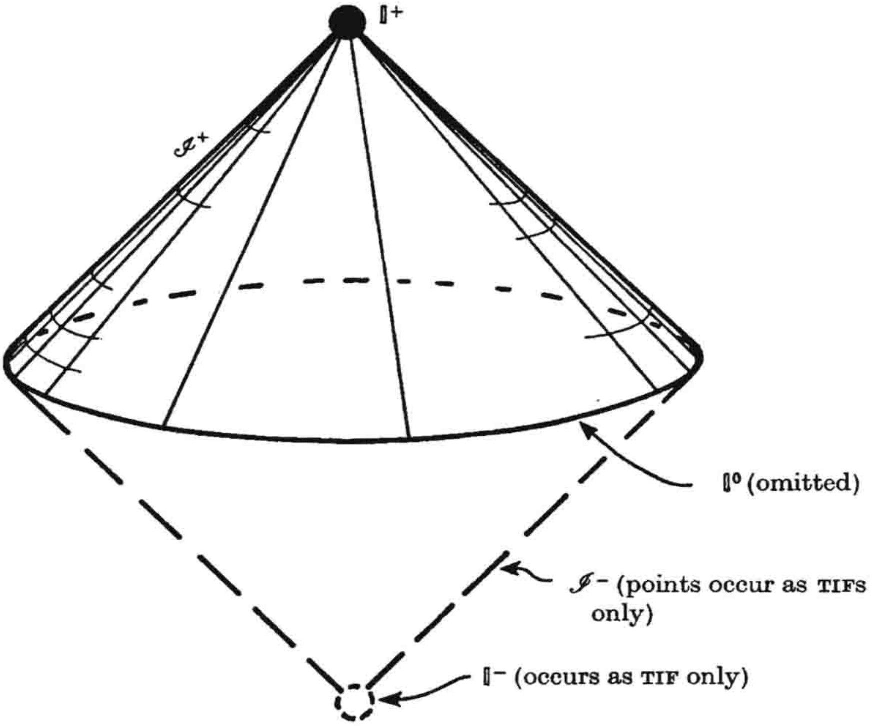

$\mathbb{I}^+$
$\mathscr{I}^+$
$\mathbb{I}^0$ (omitted)
$\mathscr{I}^-$ (points occur as TIFs only)
$\mathbb{I}^-$ (occurs as TIF only)

FIGURE 2. Minkowski space (one dimension suppressed). Adding a null surface $\mathscr{I}^+$ and a single point $\mathbb{I}^+$ to Minkowski space, we can represent each TIP as a point of $\mathscr{I}^+ \cup \{\mathbb{I}^+\}$.

For a second example, we take the Schwarzschild solution, as extended by Synge (1950) and Kruskal (1960). Here some of the ideal points represent singular points. The TIPs and TIFs of this space–time are shown in figure 3. Except for the six vertices, each point of the figure represents an entire 2-sphere: either of points of the manifold (when the point in question is an interior point of the figure) or of ideal points (when it is a boundary point). Each of the points a, b, c, d, represents a single TIP or TIF (timelike infinity). The hypersurfaces labelled '$r = \infty$' represent TIPs or TIFs at conformal infinity, while '$r = 0$' represents the set of singular points. In fact, each of the '$r = 0$' TIPs [resp. TIFs] is the past [resp. future] of some incomplete timelike geodesic.

There is an alternative, and in some respects more satisfactory, way of characterizing the IPs of a space–time. A set P in M will be called a *past-set* if P coincides with its own past, or – what is the same thing – if P is the past of *some* set. A past-set will be called *indecomposable* if it is not empty and cannot be expressed as a union of two proper subsets which are themselves past-sets. We shall now prove that the past-sets indecomposable in this sense are precisely the IPs already defined (as the pasts of timelike curves): a result anticipated by our notation.

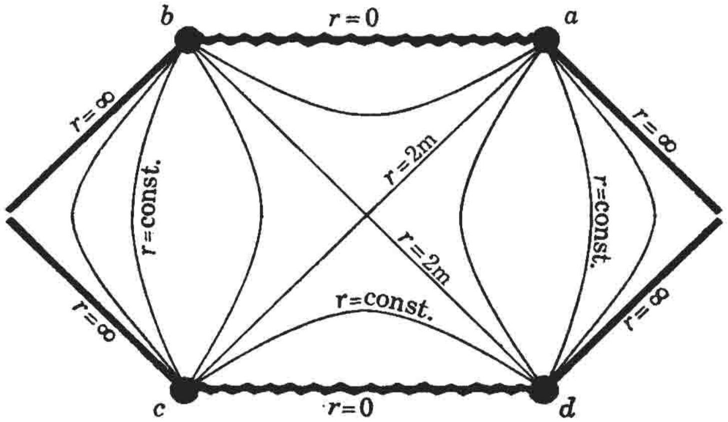


FIGURE 3. The TIPs and TIFs of the (extended) Schwarzschild solution. Each interior point of the figure represents a 2-dimensional sphere. Radial null geodesics are represented in the figure by straight lines at 45° to the vertical. The TIPs [resp. TIFs] can be represented by a three-dimensional surface, partly at 'future [resp. past] infinity' ($r = \infty$) and partly at the 'future [resp. past] singularity' ($r = 0$): a, b, c, d represent single TIPs or TIFs.

THEOREM (2.1). *A subset of a space–time is of the form $I^-[\gamma]$ for some timelike curve γ, if and only if it is an indecomposable past-set.*

(Note that the foregoing definitions, this theorem, and the proof which follows, each have a *dual*, obtained by interchanging 'past' and 'future', 'I^-' and 'I^+', throughout. We shall normally omit the duals of our statements.)

Proof of (2.1). Let γ be a timelike curve, and suppose that $I^-[\gamma] = Q \cup R$, where Q and R are past-sets. We must show that one of the two sets, Q or R, contains the other. Suppose, on the contrary, that we can find points $q \in Q - R$ and $r \in R - Q$. Since q and r lie in the past of γ, there exist points q' and r' on γ such that $q \in I^-(q')$ and $r \in I^-(r')$. But γ is timelike, and so one of the points, q' or r', must lie in the past of (or coincide with) the other. Therefore, whichever of the sets, Q or R, contains the futuremost of q' and r' must also contain both q and r, which contradicts the construction of q and r. We conclude that $I^-[\gamma]$ is indecomposable.

To prove the converse, let P be any indecomposable past-set. We must find some timelike curve γ such that $P = I^-[\gamma]$. Let p be any point of P. Then, since P is a past-set,

$$P = I^-[P \cap I^+(p)] \cup I^-[P - I^+(p)].$$

But P is indecomposable: so either $P = I^-[P \cap I^+(p)]$ or $P = I^-[P - I^+(p)]$. Since $p \notin I^-[P - I^+(p)]$, we conclude that $P = I^-[P \cap I^+(p)]$. It follows that P has this property: *given any two points in P, there exists a point in P which is to the future of both.* Now choose a countable dense family $p_0, p_1, \ldots$ of points of P. Choose a point q_0 in P to the future of p_0. Since q_0 and p_1 are in P, one can choose a point q_1 in P to the future of both; since q_1 and p_2 are in P, choose q_2 in P to the future of both of these, and so on. Since each point of the sequence $q_0, q_1, \ldots$ lies in the past of its successor, one can construct a timelike curve γ in P connecting all the points of the sequence. For each point p of P, the set $P \cap I^+(p)$ is open in P (and non-void), and must therefore – since the p_n are dense in P – contain p_k for some k. But $p_k \in I^-(q_k)$, whence $p \in I^-[\gamma]$. We have shown that every point of P lies in the past of γ, and so – since γ is contained in the past-set P – we must have $P = I^-[\gamma]$.

Remark. It is easily verified that, if P is any non-void past-set, the property italicized in the foregoing proof is in fact completely equivalent to the indecomposability of P.

We can extend our discussion to curves which are *causal* – that is, everywhere timelike or null. It will be helpful to note the following result which holds in any space-time. (We agree to call a curve with past end-point x and future end-point y a curve *from x to y*.)

(2.2). *If there exists a timelike curve from p to q and a causal curve from q to r, there exists a timelike curve from p to r.*

This result, together with the Remark above, implies that $I^-[\gamma]$ is still an indecomposable past-set when γ is a causal curve. Moreover, we have:

THEOREM (2.3). *Let γ be a causal curve. Then $I^-[\gamma]$ is a PIP, $I^-(p)$, if and only if γ has a future end-point p. Consequently $I^-[\gamma]$ is a TIP if and only if γ is a future-endless.*

It will be convenient to prove this by using a slight variant of lemma I (p. 214) of Penrose (1968) (cf. also Penrose 1972):

LEMMA (2.4) *Let S be a set and p a point for which $I^-(p) \subset I^-[S]$. If p possesses a neighbourhood N such that $I^-[S-N] = I^-[S]$, then there exists a causal geodesic from p to some point q $(\neq p)$ for which $I^-(q) \subset I^-[S]$.*

Proof of lemma. Without loss of generality we may assume that the (open) neighbourhood N of p has a compact, simply convex closure $\bar{N}$ such that $I^-[S-\bar{N}] = I^-[S]$. Let (x_n) be a sequence of points in $N \cap I^-(p)$ converging to p. For each n,

$$x_n \in I^-[S-\bar{N}];$$

so we can choose a timelike curve δ_n from x_n to some point $z_n \in S-\bar{N}$. Define y_n to be the first point at which δ_n intersects the (compact) boundary $\dot{N}$ of N: then the sequence (y_n) has an accumulation point $q \in \dot{N}$. As $p \notin \dot{N}$, certainly $q \neq p$. Since (x_n) converges to p and a subsequence of (y_n) to q, and since the set $\bar{N}$ contains timelike geodesics from x_n to y_n for every n, $\bar{N}$ (being compact and simply convex) also contains a causal geodesic from p to q. Let $x \in I^-(q)$. Then the future of x is a neighbourhood of q and so contains y_m for some m. Hence $x \in I^-(z_m) \subset I^-[S]$.

Proof of (2.3). Let $I^-[\gamma] = I^-(p)$, and suppose that p is not a future end-point of γ and therefore has a neighbourhood N for which $I^-[\gamma - N] = I^-[\gamma]$. By lemma (2.4) and (2.2), there exists a point q ($\neq p$) such that $I^-(p) \subset I^-(q) \subset I^-[\gamma]$, in contradiction of the fact that $p \neq q$ implies $I^-(p) \neq I^-(q)$. The converse proposition is a corollary of (2.2).

An important role in the proofs of theorems on singularities is played by the sets which are the boundaries of past-sets.[4] Of particular significance are those boundaries (sometimes referred to as 'null horizons') which are *ruled by future-endless null geodesics* – that is, boundaries every point of which is the past end-point of some future-endless null geodesic contained in the boundary. The following theorem – whose proof rests on (2.3) – provides a simple characterization of such boundaries in terms of TIPS.

THEOREM (2.5). *The boundary $\dot{P}$ of a past-set is ruled by future-endless null geodesics if and only if P can be expressed as a union of* TIPS.

Proof. Suppose $\dot{P}$ is ruled by future-endless null geodesics. Let $p \in P$ and choose a future-endless timelike curve γ through p. If $\gamma \subset P$, then $p \in I^-[\gamma] \subset P$. If $\gamma \not\subset P$, then $p \in I^-[\eta] \subset P$, where η is a future-endless null geodesic contained in $\dot{P}$ whose past end-point is the intersection of γ with $\dot{P}$. In either case, some subset of P is a TIP containing p.

Conversely, suppose P is a union of TIPS – say $P = \cup_i I^-[\gamma_i]$, where the γ_i are future-endless timelike curves. Let $x \in \dot{P}$ and choose a neighbourhood N of x small enough to ensure that every timelike curve in N has a future end-point (either in N or in its closure). Then $I^-[\gamma_i - N]$ must coincide with $I^-[\gamma_i]$ for each i, and consequently

$$P = \cup I^-[\gamma_i - N] = I^-[\cup \gamma_i - N] \subset I^-[P - N],$$

so that, by lemma (2.4), there exists a causal geodesic ν from x to some point z ($\neq x$) for which $I^-(z) \subset P$. Since x lies in $\dot{P}$, so does ν, and since $\dot{P}$ is achronal,[5] ν must be a null geodesic in $\dot{P}$ with past end-point x. Let μ be a null geodesic which is maximal in these respects. If μ had a future end-point x^*, the preceding argument could be repeated to construct a null geodesic μ^* in $\dot{P}$ with past end-point x^*, which (since $\dot{P}$ is achronal) would necessarily be a continuation of μ. It follows that μ is future-endless.

Space–times which are asymptotically simple (e.g. Minkowski space) are particularly easy to analyse in relation to their TIP and TIF structures. For our purpose, an *asymptotically simple* space–time is one satisfying the following conditions. There exists a set $\mathscr{I}$ which can be adjoined to M in such a way that the union $\bar{M} = M \cup \mathscr{I}$ has the structure of a manifold-with-boundary, the boundary constituting the set $\mathscr{I}$; each point of $\mathscr{I}$ is either the past or future end-point of some null geodesic in M; and every maximally extended null geodesic in M has precisely two end-points (one past, one future) on $\mathscr{I}$. We require, moreover, that $\bar{M}$ can be given a conformal structure which induces the correct conformal structure on M.

The points of $\mathscr{I}$ can be divided into two (not necessarily disjoint) classes, $\mathscr{I}^+$

and $\mathscr{I}^-$, where $\mathscr{I}^+$ consists of future, $\mathscr{I}^-$ of past end-points of null geodesics in M. The points of $\mathscr{I}$ at which $\mathscr{I}$ is timelike belong to both classes (for example, point a in figure 4), those at which $\mathscr{I}$ is spacelike to just one (points b and c).

The points at which $\mathscr{I}$ is null normally do not belong to more than one class (points d and e in figure 4), but exceptions can occur (point f). In fact, it will be convenient to impose a further condition on M (Penrose 1965*b*) which actually rules out all such exceptional situations and which is automatically satisfied if $\mathscr{I}$ is either everywhere null or everywhere spacelike:

(2.6). *Any null geodesic in* $\bar{M}$ *which at some point is tangent to* $\mathscr{I}$ *lies entirely in* $\mathscr{I}$.

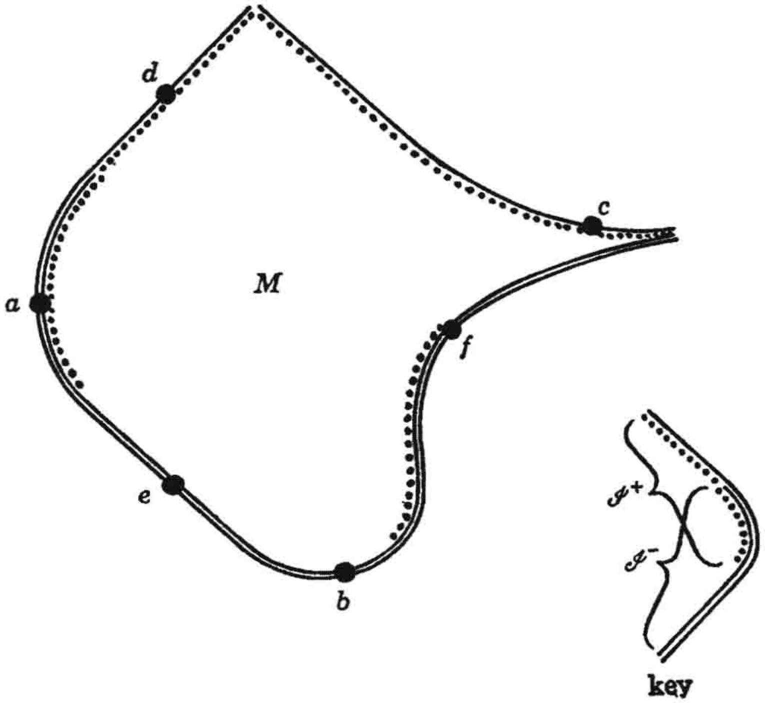


FIGURE 4. An asymptotically simple space–time (two spatial dimensions suppressed). The boundary curve represents $\mathscr{I}$. One portion of $\mathscr{I}$ (including points a, c, d and f) is $\mathscr{I}^+$, and another portion (including a, b, e and f) is $\mathscr{I}^-$.

Under this condition, we shall now show that the points of $\mathscr{I}^+$ and $\mathscr{I}^-$ can be respectively identified with the TIPs and the TIFs of M.

Let us agree to write '$I^-(p)$' for the set of all points in M which are the past end-points of timelike curves in M with future end-point p: not only, as before, when $p \in M$, but also now when $p \in \mathscr{I}^+$.

THEOREM (2.7). *For each point* p *in* $\mathscr{I}^+$, *the set* $I^-(p)$ *is a* TIP *of* M: *and conversely each* TIP (*except* M *itself if it is a* TIP) *is of the form* $I^-(p)$ *for one and only one* p *in* $\mathscr{I}^+$.

Proof. The first statement is clear: for, given $p \in \mathscr{I}^+$, there exists a null geodesic μ in M whose future end-point is p. Since μ is future-endless in M, the set $I^-(p) = I^-[\mu]$ is a TIP.

To prove the second statement, let P be a TIP distinct from M. Since P is a TIP, it is the past of some timelike curve γ in M. Since $P \neq M$, the boundary $\dot{P}$ of P in M contains (theorem (2.5)) a null geodesic ν whose future end-point in $\mathscr{I}^+$ we call p. If p is a future end-point of γ, then $P = I^-(p)$; therefore let us suppose it is not, so that $P = I^-[\gamma - N]$ for some neighbourhood N of p in $\bar{M}$. We can extend M and its conformal structure beyond the boundary $\mathscr{I}$ in a neighbourhood of p so as to form a space–time M^* with a metric consistent with this conformal structure.

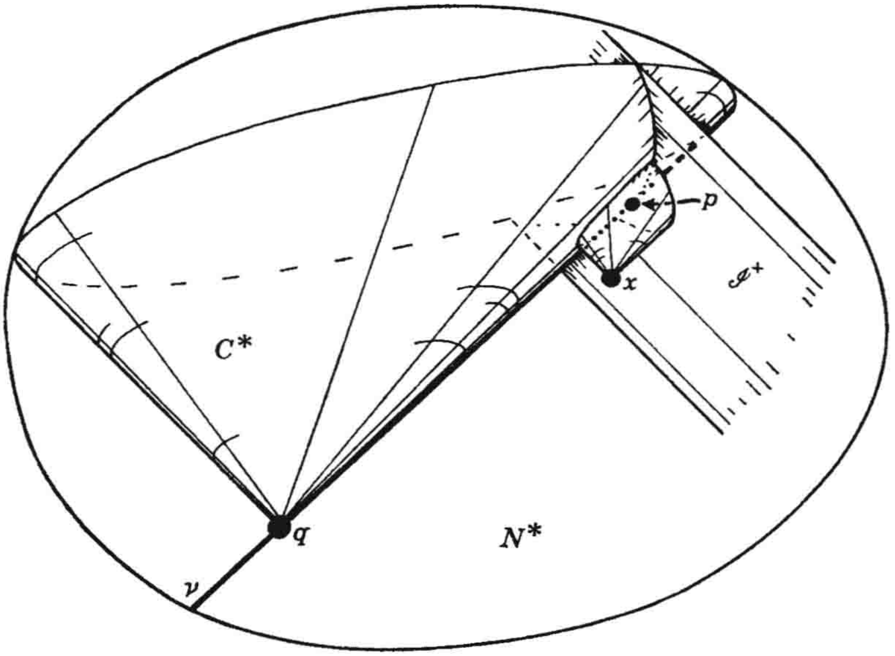


FIGURE 5. Illustration for the proof of theorem (2.7).

Let N^* be a simply convex neighbourhood in M^* of p; choose a point q in $N^* \cap \nu$; and let C^* denote the union of the null geodesics in N^* from q (figure 5). Since C^* is a null hypersurface (non-singular since N^* is simply convex), it intersects $\mathscr{I}^+$ in a set which is, by (2.6), spacelike and smooth. This, however, implies that the local geometry at p is essentially that of the configuration of tangent hyperplanes to C^* and $\mathscr{I}^+$ at p, in the Minkowski tangent-space at p. Consequently, any timelike curve from a point x chosen in P just to the past of p can leave the neighbourhood $N \cap N^*$ of p in N only by intersecting one of the hypersurfaces C^* and $\mathscr{I}^+$ and therefore (since P is disjoint from both) leaving P: in contradiction of the fact that, since $P = I^-[\gamma - N]$, some timelike curve – necessarily contained in P – links x to a point of γ not in N.

We have shown that $P = I^-(p)$. To conclude the proof, suppose that $P = I^-(p')$, where p' is a point of $\mathscr{I}^+$ distinct from p. By repeating the argument just used, with $\{p'\}$ replacing γ, we again obtain a contradiction.

Remark. The submanifold

$$\{(t, x) : t^3 + x^2 < 0\}$$

of two-dimensional Minkowski space demonstrates that, without the condition (2.6), theorem (2.7) can fail; for the TIP determined by the curve $x = 0$, $t < 0$ is not the past of any point of $\mathscr{I}$. Even to demand that $\mathscr{I}$ should be everywhere timelike is not enough: consider the submanifold

$$\{(t, x, y): y < x^2 + y^2 < 2y\}$$

of three-dimensional Minkowski space and the TIP determined by the curve

$$4x = 3\sin t, \quad 4y = 3 - 3\cos t, \quad 0 < t < 2\pi.$$

To conclude this section, we return to the case of a general space–time M and shall formulate the notion of 'domain of dependence' in terms of indecomposable sets. One advantage of this formulation will be that it can be utilized to define domains of dependence in the more general context of causal spaces (see § 3).

Let S be an achronal set in M. (In physical problems one normally takes S to be a spacelike or null hypersurface.) The *future domain of dependence* $D^+(S)$ of the set S is defined (cf. Penrose 1968, Geroch 1970*a*) as the collection of all points p of M such that every past-endless timelike curve through p intersects S.

To characterize $D^+(S)$ in terms of IFs, observe first that the set $I^+[S]$, being open in M, is a space–time in its own right. For each s in S, the set $I^+(s)$ is clearly a TIF in the space–time $I^+[S]$. However, there may be TIFs which are not of this form; and these represent precisely the regions which must be deleted from $S \cup I^+[S]$ to obtain $D^+(S)$:

THEOREM (2.8). *Let S be an achronal set in M. Then* $D^+(S) = S \cup I^+[S] - K$, *where K is the union of all* TIFs *of the space–time* $I^+[S]$ *which are not of the form* $I^+(s)$ *with s in S.*

Proof. Since $S \subset D^+(S) \subset S \cup I^+[S]$, we need only show that $I^+[S] - D^+(S) = K$.

Let $p \in I^+[S] - D^+(S)$. Then there is some past-endless timelike curve γ through p which does not intersect S. This curve may remain in $I^+[S]$ or it may not: but in either case $I^+[\gamma \cap I^+[S]]$ is a TIF of the space–time $I^+[S]$. This TIF cannot be of the form $I^+(s)$ with $s \in S$; for if it were, γ would intersect S at s (theorem (2.3)). It follows that $p \in K$.

Conversely, let $p \in F$, where F is a TIF of the space–time $I^+[S]$ which cannot be written in the form $I^+(s)$ for any s in S. By (2.3) we can write $F = I^+[\mu]$, where μ is a timelike curve in $I^+[S]$ which passes through p and is past-endless in $I^+[S]$. If μ is past-endless in M, then $p \notin D^+(S)$. If, however, μ has a past end-point q in M, then q belongs to neither S nor $I^+[S]$, so that a past-endless timelike curve through q and p (which exists, since $q \in I^-(p)$) will not intersect S; and we again conclude that $p \notin D^+(S)$.

Remark 1. That Cauchy horizons are effectively unions of null geodesics (Hawking 1966*a*, 1967*a*; cf. also Penrose 1968, Geroch 1970*a*, Hawking & Penrose 1970) is an immediate consequence[6] of theorems (2.5) and (2.8).

Remark 2. A *Cauchy hypersurface* in M – i.e. an achronal set S such that

$$M = D^+(S) \cup D^-(S)$$

– may be defined quite simply in terms of indecomposable sets. An achronal set S is a Cauchy hypersurface for a space–time M if and only if $M = S \cup I^+[S] \cup I^-[S]$ and every TIP and every TIF of M intersects S.

3. IDEAL POINTS OF CAUSAL SPACES

We shall now extend certain of the ideas of § 2 to causal spaces. This will serve to illustrate how far our constructions are actually independent of the details of the differentiable structure of space–times.

A *causal space* (Kronheimer & Penrose 1967) is a set X equipped with two *causal relations*, $\prec$ and $\ll$, between the points of X, such that (1) $\prec$ is a reflexive (i.e. $x \prec y$ and $y \prec x$ if and only if $x = y$) partial ordering, (2) $\ll$ is an anti-reflexive (i.e. $x \ll x$ for no x) partial ordering, (3) $x \ll y$ implies $x \prec y$, and (4) $x \ll y \prec z$ and $x \prec y \ll z$ each imply $x \ll z$. Any subset X' of X can itself be regarded as a causal space by restricting $\prec$ and $\ll$ to the points of X'. The relations $\prec$ and $\ll$ are called, respectively, the *causality* and the *chronology*. Our discussion of ideal points will be based almost entirely on the chronology.

The concept of a causal space generalizes and abstracts certain aspects of the structure of a space–time having no self-intersecting causal curves. Given that we admit only space-times which are past- and future-distinguishing, every space–time has the structure of a causal space:[7] we put $x \ll y$ [resp. $x \prec y$] whenever there exists a timelike [resp. causal] curve with past end-point x and future end-point y [and, additionally, put $x \prec y$ if $x = y$].

The (*chronological*) *past* of a point x in the causal space X is the set

$$I^-(x) = \{y : y \ll x\},$$

and X will be *past-distinguishing* if distinct points have distinct pasts. The *past of a subset* S of X is the set $I^-[S] = \cup_{x \in S} I^-(x)$, and S is called a *past-set* if $S = I^-[S]$. All these definitions coincide with those of § 2 when X is a space–time. (Note that we shall continue to omit *dual* statements, obtained by interchanging 'past' and 'future'.)

As we have seen, in any space–time the past of a set is necessarily a past-set; but this is not true for all causal spaces without further conditions. The natural condition to impose upon a causal space in examining the structure of its past-sets turns out to be a weakened form of the 'fullness' considered in Kronheimer & Penrose (1967). A causal space will be called *past semi-full* if each point has a chronological predecessor and if, whenever $p_1 \ll r$ and $p_2 \ll r$, there exists a point $q \ll r$ such that $p_1 \ll q$ and $p_2 \ll q$. (A causal space is *full* if it is both past and future semi-full. Clearly any space–time is full.) The notion of past semi-fullness may be reformulated as follows : the causal space X is past semi-full if and only if, for each

As far as we have been able to ascertain, our construction yields the obviously 'reasonable' completion whenever there is one. We do not wish to imply, however, that the $\bar{M}$ we suggest is the only possible completion, nor that it is necessarily the best or the simplest. In fact, several alternative possibilities had suggested themselves previously, but each was found to yield unsatisfactory answers in certain fairly simple cases. The definition for $\bar{M}$ that we give here is the only survivor, but the reader should not be discouraged from suggesting alternatives. We would recommend, however, that any such alternative be checked against the examples discussed below.

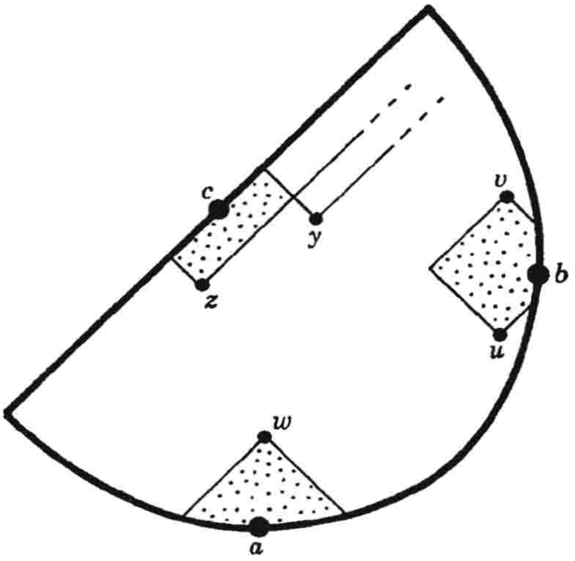


FIGURE 6. An asymptotically simple space–time (two spatial dimensions suppressed). Appropriate neighbourhoods of point a, where $\mathscr{I}$ is spacelike, and point b, where $\mathscr{I}$ is timelike, can be constructed in the Alexandrov topology. This is not so for the point c, where $\mathscr{I}$ is null. There is, however, a base of neighbourhoods of c of the form $I^+(z) - \overline{I^+(y)}$.

One might conjecture that some form of 'Alexandrov topology' (cf. Kronheimer & Penrose 1967) would be appropriate for $\hat{M}$. In the Alexandrov topology for a full causal space X, there is a base for the neighbourhoods of any point x consisting of all sets of the form $I^+(u) \cap I^-(v)$ with $u \ll x \ll v$: equivalently, the Alexandrov topology is the coarsest topology in which all PIPs and PIFs are open. However, we cannot simply carry this topology over to $\hat{X}$ (using, say, the chronological relation $\ll_{\mathrm{IP}}$ of the last section) since, even in Minkowski space, we would not obtain sufficiently many open sets in the IP-space. Indeed, it is intuitively clear from figure 6 that if M is an asymptotically simple space–time, then, although appropriate neighbourhoods of points at which $\mathscr{I}$ is spacelike or timelike can be constructed by means of sets such as $I^+_{\mathrm{IP}}(U)$, $I^-_{\mathrm{IP}}(V)$, and $I^+_{\mathrm{IP}}(U) \cap I^-_{\mathrm{IP}}(V)$ – where $U = I^-(u)$, $V = I^-(v)$ – this is not in general so for points at which $\mathscr{I}$ is null. The problem, in the case of a null $\mathscr{I}^+$, is that there are no IPs to 'cut off' the neighbourhood from above at points such as c. Since the IPs will not do, one might envisage using for this a past semi-full causal space in which the IP $I^-[Q_1 \times \{\omega\}]$ is generated by no chronological chain.) However, the argument used to prove theorem (2.1) shows that an IP P will be the past of a chronological chain contained in P if P is the past of some countable set contained in P – i.e. if P is separable in some topology in which $P \cap I^+(x)$ is open for each x. (The last example shows that the phrase 'contained in P' cannot be omitted.) We use this condition in theorem (3.6).

As before, we shall divide the IPs of X into PIPs and TIPs, according as they are or are not of the form $I^-(x)$ for some $x \in X$.

To specify what we mean by a domain of dependence, lacking the concept of curves, we again use a theorem already proved in the context of space–times – in this case (the dual of) theorem (2.8) – as the basis for a definition. Let S be an *achronal* set – i.e. a set no two points of which are chronologically related. Recall that $S^{(-)}$, the union of S with its past, is a past semi-full causal space (lemma (3.1)). We now define the *past domain of dependence* $D^-(S)$ of S, to be the set of all points of the space $S^{(-)}$ which do not belong to any TIP of that space.

It turns out that this definition allows us to characterize the domain of dependence in terms of chronological chains. Let us agree to call q a *future end-point* of the causal chain C in X if $I^-(q) = I^-[C]$ – a definition suggested by theorem (2.3) – and let us assume that X is past-distinguishing. If C is a chronological chain and has a future end-point q, then q is unique and $C \cup \{q\}$ is again a chronological chain. Consequently, any chronological chain which has a future end-point in $S^{(-)}$ can be extended to one which intersects S by adding at most one point to it.

THEOREM (3.6). *Suppose that every IP in X contains a chronological chain which generates it. Let S be an achronal set and p a point of $S^{(-)}$. Then $p \in D^-(S)$ if and only if every chronological chain in $S^{(-)}$ whose first point is p has a future end-point in $S^{(-)}$.*

Proof. Suppose $p \notin D^-(S)$. Then p belongs to some TIP P of $S^{(-)}$. Since P is certainly an IP of X, it is generated by some chronological chain $G \subset P \subset S^{(-)}$. It follows from lemma (3.3) that $(I^+(p) \cap G) \cup \{p\}$ is a chronological chain, beginning at p, which still generates P, and which (since P is not of the form $I^-(q)$ for any q in $S^{(-)}$) can therefore have no future end-point in $S^{(-)}$.

Conversely, suppose $p \in D^-(S)$; and let C be a chronological chain in $S^{(-)}$ whose first point is p. Then $I^-[C]$ is an IP (theorem (3.5)); and hence either $I^-[C] = I^-(q)$ for some $q \in S^{(-)}$, or else $I^-[C]$ is a TIP of $S^{(-)}$ and does not therefore contain the point p. In the first case, q is a future end-point of C; in the second, p must be the only member of C and therefore its future end-point.

We shall denote by $\hat{X}$ the collection of all IPs of the past semi-full causal space X. The set $\hat{X}$ may itself be given a causal structure by writing (for each $P, Q \in \hat{X}$):

$$P \prec_{\mathrm{IP}} Q \quad \text{if} \quad P \subset Q;$$
$$P \ll_{\mathrm{IP}} Q \quad \text{if} \quad P \subset I^-(q) \quad \text{for some} \quad q \in Q.$$

The verification that these relations define a causal structure on $\hat{X}$ is trivial. The causal space $\hat{X}$ will be called the *IP-space* of X. We let $I^-_{\mathrm{IP}}(Q)$ denote the collection of all IPs P of X for which $P \ll_{\mathrm{IP}} Q$.

In order to discuss IP-spaces, it will be convenient to call any causal space *future-complete* if it satisfies the following three conditions:

FC 1. The space is past semi-full.

FC 2. Every IP of the space is a PIP.

FC 3. If $I^-(x) \subset I^-(y)$, then[9] $x \prec y$.

Note that ***FC 3*** implies that the causal space is past-distinguishing.

The fundamental property of IP-spaces is

THEOREM (3.7). *Let X be a past semi-full causal space. Then $\hat{X}$ is future-complete.*

Proof. ***FC 1.*** Let $P \in \hat{X}$. Since P is not empty, neither is $I_{\mathrm{IP}}^-(P)$. Let Q and Q' be members of $I_{\mathrm{IP}}^-(P)$. Then there exist $q, q' \in P$ such that $Q \subset I^-(q)$ and $Q' \subset I^-(q')$. Since P is an IP, it contains points r and s such that $q \ll s$, $q' \ll s$, and $s \ll r$ (lemma (3.4)). Writing $S = I^-(s)$, it follows that $Q \ll_{\mathrm{IP}} S$, $Q' \ll_{\mathrm{IP}} S$, and $S \ll_{\mathrm{IP}} P$: that is, $\hat{X}$ is past semi-full.

FC 2. Let $\mathscr{P}$ be an IP of $\hat{X}$, and thus a collection of IPs of X. Then $P = \bigcup_{Q \in \mathscr{P}} Q$ is a past-set of X. We first show that P is an IP. Since $\mathscr{P}$ is not empty, neither is P. Let q and q' be points of P. Then there must be IPs Q and Q' belonging to the collection $\mathscr{P}$ such that $q \in Q$ and $q' \in Q'$. But $\mathscr{P}$ is an IP of $\hat{X}$, and so we can find a member R of $\mathscr{P}$ such that $Q \ll_{\mathrm{IP}} R$ and $Q' \ll_{IP} R$. That is to say, there must be points r and r' in R such that $Q \subset I^-(r)$ and $Q' \subset I^-(r')$. But R itself is an IP of X, and so contains a point s succeeding both r and r', whence $q \ll s$ and $q' \ll s$. Since $R \subset P$, we have shown that P contains a successor of both q and q' and, therefore, that P is an IP of X. We conclude the proof by showing that $\mathscr{P} = I_{\mathrm{IP}}^-(P)$. Let $U \in \mathscr{P}$. Then, since $\mathscr{P}$ is a past-set of $\hat{X}$, there exists a $V \in \mathscr{P}$ such that $U \ll_{\mathrm{IP}} V$. But $V \subset P$, so $U \ll_{\mathrm{IP}} P$; i.e. $U \in I_{\mathrm{IP}}^-(P)$. Conversely, suppose $W \in I_{\mathrm{IP}}^-(P)$. Then $W \subset I^-(p)$ for some point p of P. Choosing a member Q of $\mathscr{P}$ containing p, we have $W \ll_{\mathrm{IP}} Q \in \mathscr{P}$. Therefore, since $\mathscr{P}$ is a past-set in $\hat{X}$, $W \in \mathscr{P}$.

FC 3. Let $I_{\mathrm{IP}}^-(P) \subset I_{\mathrm{IP}}^-(Q)$. For each point p of P, we can find $x, y \in P$ such that $p \ll x \ll y$. Consider the IP $I^-(x)$. Since $I^-(x) \ll_{\mathrm{IP}} P$ and $I_{\mathrm{IP}}^-(P) \subset I_{\mathrm{IP}}^-(Q)$, we must have $I^-(x) \ll_{\mathrm{IP}} Q$. Consequently, $p \in I^-(x) \subset Q$. We have shown that each point of P belongs to Q: i.e. that $P \prec_{\mathrm{IP}} Q$.

Recall that the IPs of the causal space X are divided into two classes: the PIPs and TIPs. With each point x of X we may associate the PIP $I^-(x)$, thus defining a mapping, conveniently denoted by I^-, from X into $\hat{X}$. It is clear that, if $p \prec q$, then $I^-(p) \prec_{\mathrm{IP}} I^-(q)$: that is, the mapping I^- *preserves the causality*. Similarly, I^- *preserves the chronology*. The range of the mapping I^- is just the collection, $\hat{X}_{\mathrm{PIP}}$, of all PIPs of X. If the mapping is one-to-one, it has an inverse from $\hat{X}_{\mathrm{PIP}}$ onto X. If, moreover, this inverse mapping happens itself to preserve both causality and chronology, then $\hat{X}_{\mathrm{PIP}}$ has not only essentially the same points but also essentially the same causal structure as X. More precisely, I^- is a *causal isomorphism* from X to $\hat{X}_{\mathrm{PIP}}$, and the spaces X and $\hat{X}_{\mathrm{PIP}}$ are *causally isomorphic*.

We summarize various possibilities in the following theorem:

THEOREM (3.8). *Let X be a past semi-full causal space. Then*

(1) *I^- is one-to-one if and only if X is past-distinguishing;*

(2) *I^- maps X onto $\hat{X}$ if and only if every* IP *of X is a* PIP*;*

(3) *I^- is a causal isomorphism from X onto $\hat{X}_{\mathrm{PIP}}$ if and only if X satisfies* ***FC3****;*

(4) *I^- is a causal isomorphism from X onto $\hat{X}$ if and only if X is future-complete.*

Of these statements, (1) and (2) are trivial, while (4) is an obvious consequence of (2) and (3). To establish (3), we first observe that ***FC3*** has already been noted to imply past-distinction and clearly implies also that the inverse of I^- preserves causality and chronology. The rest of (3) is an immediate consequence of the fact that the causal space $\hat{X}_{\mathrm{PIP}}$ always satisfies ***FC3***, which in turn is established by slightly re-wording the proof that $\hat{X}$ does.

From theorems (3.7) and (3.8) it follows that the process of taking IP-spaces of a past semi-full causal space need at most be performed once. Starting from a space which is not future-complete, we get a future-complete extension of that space, in the sense we have discussed; repeating the process yields essentially the same extension. However, essentially different extensions of the same general type also exist. The characterization of such extensions, independently of their method of construction, is discussed in Kronheimer (1972).

4. IDENTIFICATION AND TOPOLOGY

In the preceding two sections we have discussed how ideal points may be adjoined to any past- and future-distinguishing space–time (or full causal space) to obtain *either* $\hat{M}$ (by adjoining TIPS) *or* $\check{M}$ (by adjoining TIFS). We have not as yet suggested any method of combining these two constructions to obtain some completion $\bar{M}$ encompassing both the TIPS and TIFS. It would not be satisfactory in all cases simply to adjoin the two types of ideal points independently to M: some identifications between ideal points may also be required. For example, in the case of an asymptotically simple space–time with a timelike $\mathscr{I}$, each individual point of $\mathscr{I}$ would arise *both* as a TIP *and* as a TIF in M. On the other hand, if $\mathscr{I}$ is spacelike or null, each point of $\mathscr{I}$ arises *only* as a TIP or *only* as a TIF. We thus require some criterion for identifying ideal points, where appropriate, if we expect to obtain any reasonable definition for $\bar{M}$ in general situations.

An additional problem is that of finding a suitable topology for such an $\bar{M}$. Indeed, we have as yet given no indication as to how the topology even of $\hat{M}$ or $\check{M}$ should be defined.

In the present section we shall give a number of examples of space–times, which serve to illustrate some of the difficulties associated with combining $\hat{M}$ and $\check{M}$ into a topological space. We shall then introduce a construction which yields a Hausdorff topological space $\bar{M}$. This $\bar{M}$ – which is an identified union of $\hat{M}$ and $\check{M}$ – can be considered as a completion of M: there is a natural dense embedding of M in $\bar{M}$. The topology and identifications are obtained simultaneously in our construction.

As far as we have been able to ascertain, our construction yields the obviously 'reasonable' completion whenever there is one. We do not wish to imply, however, that the $\bar{M}$ we suggest is the only possible completion, nor that it is necessarily the best or the simplest. In fact, several alternative possibilities had suggested themselves previously, but each was found to yield unsatisfactory answers in certain fairly simple cases. The definition for $\bar{M}$ that we give here is the only survivor, but the reader should not be discouraged from suggesting alternatives. We would recommend, however, that any such alternative be checked against the examples discussed below.

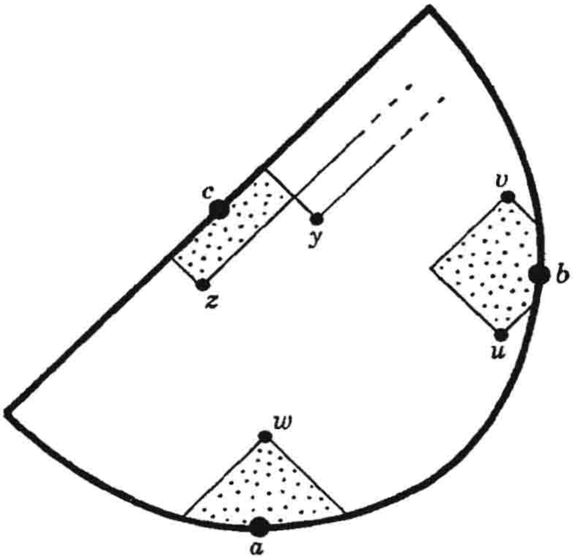


FIGURE 6. An asymptotically simple space–time (two spatial dimensions suppressed). Appropriate neighbourhoods of point a, where $\mathscr{I}$ is spacelike, and point b, where $\mathscr{I}$ is timelike, can be constructed in the Alexandrov topology. This is not so for the point c, where $\mathscr{I}$ is null. There is, however, a base of neighbourhoods of c of the form $I^+(z) - \overline{I^+(y)}$.

One might conjecture that some form of 'Alexandrov topology' (cf. Kronheimer & Penrose 1967) would be appropriate for $\hat{M}$. In the Alexandrov topology for a full causal space X, there is a base for the neighbourhoods of any point x consisting of all sets of the form $I^+(u) \cap I^-(v)$ with $u \ll x \ll v$: equivalently, the Alexandrov topology is the coarsest topology in which all PIPs and PIFs are open. However, we cannot simply carry this topology over to $\hat{X}$ (using, say, the chronological relation $\ll_{\mathrm{IP}}$ of the last section) since, even in Minkowski space, we would not obtain sufficiently many open sets in the IP-space. Indeed, it is intuitively clear from figure 6 that if M is an asymptotically simple space–time, then, although appropriate neighbourhoods of points at which $\mathscr{I}$ is spacelike or timelike can be constructed by means of sets such as $I^+_{\mathrm{IP}}(U)$, $I^-_{\mathrm{IP}}(V)$, and $I^+_{\mathrm{IP}}(U) \cap I^-_{\mathrm{IP}}(V)$ – where $U = I^-(u)$, $V = I^-(v)$ – this is not in general so for points at which $\mathscr{I}$ is null. The problem, in the case of a null $\mathscr{I}^+$, is that there are no IPs to 'cut off' the neighbourhood from above at points such as c. Since the IPs will not do, one might envisage using for this purpose what will amount to the complement of the closure of a set such as $I^+(y)$. This, however, requires a suitable notion of 'closure' in terms of the causal structure of M.

Before giving the details of the construction of $\bar{M}$, it will be worth while to point out some of the difficulties which appear to be intrinsic to the problem of defining a topology for $\hat{M}$ if we try to treat the topology in isolation from that of $\check{M}$ and from the identification problem. In intuitive terms, what we require of a topology for $\hat{M}$ is that it should tell us when we may regard an IP as 'moving continuously'. (That is, it should tell us which IPs are limits of sequences of which other IPs.) Unfortunately, it is not satisfactory simply to assert that an IP 'moves continuously' whenever it does so (in some sense) as a *point set* of M. There are several objections.

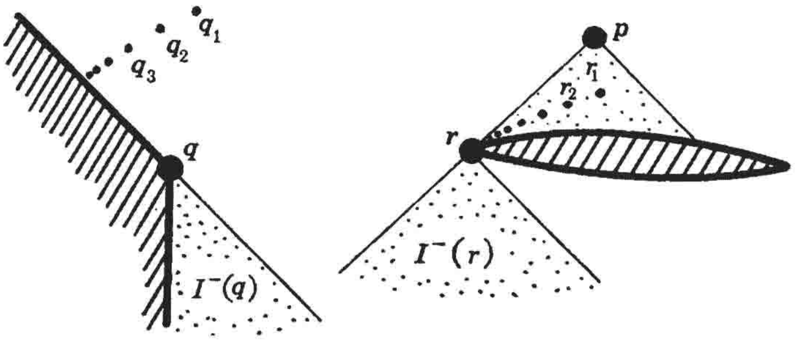


FIGURE 7. A two-dimensional space–time with two holes (striped regions). Although the $I^-(q_i)$ approach $I^-(q)$, the q_i do not appear to approach q. On the other hand, the r_i approach r, although the $I^-(r_i)$ do not appear to approach $I^-(r)$. Also the IP $I^-(p)$ changes discontinuously (to include $I^-(r)$) when p is moved slightly to the left.

Even in the case of PIPs, the set $I^-(x)$ does not move continuously with x in all space–times (consider, for instance, x approaching p from the left in figure 7), since $I^-(x)$ may encounter a 'hole' which has nothing to do with x locally. There are other more serious objections. Take the point q of figure 7, for example. Here the sets $I^-(q_i)$ converge on the TIP represented by q, although it would be difficult to think of the ideal point q as the limit of the q_i. (This example is of special interest in that the extended Reissner–Nordstrom solution (see Graves & Brill 1960, Carter 1966) exhibits this type of structure.) The problem at r is rather the opposite one. We would like to think that the points r_i are converging on the ideal point r. But the sets $I^-(r_i)$ do not in any sense have as their limit the TIP represented by r: indeed, the only reasonable limit is the empty set.

Our suggestion for circumventing these difficulties depends on the observation that if, in these particular examples, we had been concerned with IFs rather than IPs, we should have had no problems. Thus, in figure 7, the sets $I^+(q_i)$ do *not* converge on the TIF represented by q, while the sets $I^+(r_i)$ *do* converge on the TIF represented by r. If we can find some method of forcing an identification between IPs and IFs, then we may be able to use satisfactory convergence properties of the one to overcome unsatisfactory properties of the other. What we propose is that, in the first instance, the PIP and PIF defined by each point of M be identified as a single (internal) point of $\bar{M}$. Then, after introducing some open sets in accordance with a modified Alexandrov topology, we force still more identifications between elements of $\hat{M}$ and $\check{M}$ by the requirement that we should end up with the 'largest possible' Hausdorff space. In some cases this will result in identifications *within* $\hat{M}$ or $\check{M}$, as well as between these two sets. The resulting (uniquely defined) Hausdorff space will be our proposal for $\bar{M}$.

Let us now make this definition of $\bar{M}$ precise. As in § 2, we require that our space–time be both past- and future-distinguishing. In fact, we must impose the slightly stronger condition of *strong causality* (cf. Hawking 1966*a*), which states that the Alexandrov topology for M is Hausdorff. (Strong causality is the necessary and sufficient condition that the Alexandrov topology on M coincide with the ordinary manifold topology (Kronheimer & Penrose 1967).) Our definition of $\bar{M}$ will also apply when M is any full causal space, although the resulting $\bar{M}$ may be rather unsatisfactory unless conditions additional to strong causality are imposed.

We first introduce an intermediate space $M^\natural$, obtained from the union[10] $\hat{M} \cup \check{M}$ by identifying $I^+(p) \in \check{M}$ with $I^-(p) \in \hat{M}$ for each $p \in M$. If P is an element of $\hat{M}$ or $\check{M}$, we write P^* for the corresponding element of $M^\natural$. Thus, if P is a TIP or a TIF, then $P^* = Q^*$ if and only if $P = Q$; but for any $x \in M$, $(I^+(x))^* = (I^-(x))^*$.

We define our (extended Alexandrov) topology on $M^\natural$ as the coarsest topology in which, for each $A \in \check{M}$ and $B \in \hat{M}$, the four sets A^{int}, A^{ext}, B^{int}, B^{ext} are all open; where

$$A^{\text{int}} = \{P^*\colon P \in \hat{M} \text{ and } P \cap A \neq \varnothing\},$$

$$A^{\text{ext}} = \{P^*\colon P \in \hat{M} \text{ and (for all } S \subset M)\ P = I^-[S] \Rightarrow I^+[S] \not\subset A\},$$

B^{int} and B^{ext} being defined similarly, with the roles of past and future interchanged. The sets A^{int} and B^{int} are analogues, in $M^\natural$, of the $I^+(a)$ and $I^-(b)$ which define the Alexandrov topology. (For example, if $A = I^+(a)$ and $P = I^-(p)$, then $P^* \in A^{\text{int}}$ if and only if $p \in I^+(a)$ – but the definition of A^{int} also enables us to incorporate TIPs into $I^+(a)$.) The sets A^{ext} and B^{ext} are needed to obtain the appropriate topology for null $\mathscr{I}$ (cf. our requirement to construct a closure for $I^+(y)$ in figure 6).

Observe that A^{int} consists of the images in $M^\natural$ of all IPs which can be generated by subsets of A (since, if $P \in \hat{M}$ and $P \cap A \neq \varnothing$, then $I^-[P \cap A] = P$), whereas A^{ext} consists of the images in $M^\natural$ of all IPs which cannot be generated by any set which also generates a future-set contained in A. Consequently A^{int} and A^{ext} are disjoint.

We now have a topological space $M^\natural$ (which will not in general be Hausdorff). We must carry out additional identifications in $M^\natural$ to obtain $\bar{M}$. Roughly speaking, we impose the smallest possible number of identifications between points of $M^\natural$ which will suffice to make the resulting space Hausdorff (cf. Geroch 1968). (The topology of a space resulting from identifications is to be such that a set is open if and only if its inverse image under the identification map is open in the original space.) More precisely, $\bar{M}$ is the quotient space $M^\natural / R_H$, where R_H is the intersection of all equivalence relations $R \subset M^\natural \times M^\natural$ such that $M^\natural / R$ is Hausdorff.

In the case of a (strongly causal) space–time M, our definition appears to yield a plausible space $\overline{M}$. In particular, none of the identifications implied by the passage from $M^\natural$ to $\overline{M}$ occur between elements of $M^\natural$ representing points of M. In fact, there exists a natural, dense topological embedding of M in $\overline{M}$; so that we can consider $\overline{M}$ as obtained simply by adding ideal points to M. We indicate briefly how these properties of $\overline{M}$ are proved.

The natural injection ϕ: $M \rightarrow M^\natural$ (mapping each x to $(I^+(x))^* = (I^-(x))^*$) is evidently open, since $\phi[I^+(a) \cap I^-(b)]$ is open in $M^\natural$. The non-trivial part of proving ϕ is continuous (and hence an embedding) is to show that $\phi^{-1}[A^{\text{ext}}]$ and $\phi^{-1}[B^{\text{ext}}]$ are open for each $A \in \check{M}$ and $B \in \hat{M}$. The proof requires the condition

$$I^-[S] = I^-(p) \Rightarrow I^+[S] \supset I^+(p)$$

and its dual, in addition to the axioms of a full causal space. By an application of lemma (2.4) and of (2.2), all (past- and future-distinguishing) space–times are seen to satisfy this condition. The same condition is needed to prove that each non-void subset of $M^\natural$ of the form $A^{\text{int}} \cap A_1^{\text{ext}} \cap \ldots \cap A_n^{\text{ext}}$ $(A, A_1, \ldots, A_n \in \check{M})$ intersects $\phi[M^\natural]$; from which (and the dual result) it easily follows that ϕ is dense. To prove finally that the natural map $M \rightarrow \overline{M}$ is also injective (and hence a dense embedding, since ϕ is one), it suffices to show that the space $M^\natural/R$ is Hausdorff, where R is the equivalence relation which simultaneously identifies all the elements of $M^\natural$ not in $\phi[M]$. Again a condition on M additional to the causal axioms is used: for each $x \in M$, there exist a, b such that $a \ll x \ll b$ and such that there is no set S satisfying both $I^+[S] \subset I^+(a)$ and $I^-[S] \subset I^-(b)$ for which $I^+[S]$ is a TIF or $I^-[S]$ is a TIP. It is clear that this, too, is a property of any space–time.

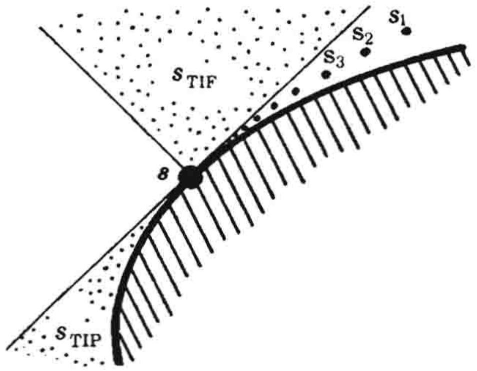


FIGURE 8. A two-dimensional space–time with a hole (striped region). The identifications used to obtain $\overline{M}$ from $M^\natural$ cause the points s_i to approach s, even though the sets $I^-(s_i)$ do not appear to approach s_{TIP}.

To ask whether $\overline{M}$ is satisfactory in other respects is a less well-defined question, since it may not be clear *a priori* what the ideal-point structure should be. One would certainly wish to reproduce the conformal boundary $\mathscr{I}$, attached to M with the correct topology, in the case of an asymptotically simple space–time. It appears, in fact, that our definition for $\overline{M}$ does this. Furthermore, $\overline{M}$ gives the expected answers in the examples of both figures 6 and 7. To see how this comes about consider the example illustrated in figure 8.

Here $M^\natural$ is such that s (the point at which $\mathscr{I}$ is null) and each point of $\mathscr{I}$ below s (i.e., where $\mathscr{I}$ is timelike) occur twice in $M^\natural$: once as a TIP and once as a TIF. Each point of $\mathscr{I}$ above s (where $\mathscr{I}$ is spacelike) occurs only once (as a TIF). Now each point q of $\mathscr{I}$, in its role as a TIF (q_{TIF}) possesses an adequate neighbourhood of the form B^{int} or $B^{\mathrm{int}} \cap C^{\mathrm{ext}}$, with $B, C \in \hat{M}$. The same conclusion holds for q in its role as a TIP (q_{TIP}), provided $\mathscr{I}$ is timelike at q. However, the neighbourhoods of s_{TIP} in $M^\natural$ look unreasonable, since the points s_i, although appearing to approach s, do not in fact enter every neighbourhood of s_{TIP}. But we have yet to carry out the identifications leading to $\bar{M}$. We observe that, if q is any point of $\mathscr{I}$ at which $\mathscr{I}$ is timelike or null, every neighbourhood of q_{TIP} intersects every neighbourhood of q_{TIF} (in M); therefore, in the passage to $\bar{M}$, the points q_{TIP} and q_{TIF} will be identified. This applies, in particular, to s. The open sets in $\bar{M}$ must have inverse images open in $M^\natural$. Hence the 'unreasonable' neighbourhoods of s_{TIP} are eliminated in passing to $\bar{M}$. No further identifications are forced on us, since we have arrived at a Hausdorff space. The resulting $\bar{M}$ in fact agrees with the intuitive topology of figure 8.

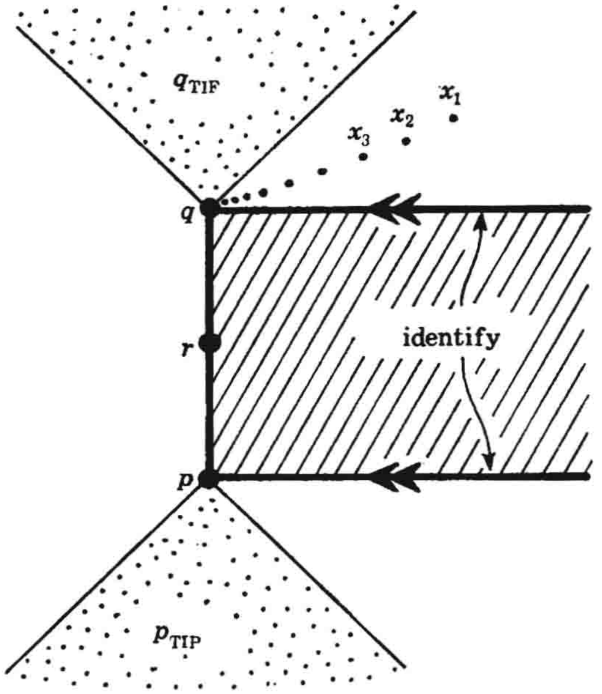


FIGURE 9. Two-dimensional Minkowski space with a region (striped) removed and with the edges indicated identified. With any reasonable identification procedure we would expect $p = q$; yet with any reasonable causal structure we would expect $p \ll r$ and $r \ll q$.

Finally we consider the question, whether one would expect to find, in general, a reasonable causal structure on $\bar{M}$. One obstacle to this is illustrated in figure 9. Because p_{TIP} and q_{TIF} are both limits of the sequence x_i (approaching from the right), the ideal points p and q will be identified in $\bar{M}$. However, there is an entire line segment of ideal points from p to q, including in particular r; and with a causal structure on $\bar{M}$, we should expect both $p \ll r$ and $r \ll q$. But if, in addition, $p = q$, this would contradict the axioms for a causal space.

It is not intended that these examples should be thought of as in any sense realistic as regards, say, space–time singularities. It may be that Einstein's equations or energy conditions imply severe restrictions on the nature of the singular boundaries that can occur. Nevertheless, since so little is known at present about the details of the structure of space–time singularities, one would like to have techniques that can – in principle, at least – cope with virtually any type of situation consistent with the requirements of causality.

NOTES

[1] For further discussion of this point see, for example, Geroch (1968).

[2] By a *space–time* we normally mean a smooth connected Hausdorff 4-manifold with a smooth metric of signature $(+, -, -, -)$, which is time-oriented (i.e. the two null half-cones at each point are continuously labelled 'past' and 'future'). However, in many of our examples we shall, for simplicity, consider two- or three-dimensional space–times of signature $(+, -)$ or $(+, -, -)$.

[3] We use '*curve*' to mean either a map with non-zero derivative of the real line into the manifold or else the image of such a map. An *end-point* of a curve is a point of the manifold every neighbourhood of which the curve enters and remains in as the parameter tends to ∞ or to $-\infty$. For a timelike curve (and for a causal curve [see p. 551]) the terms *past* and *future end-point* have their obvious meanings. '*Past-endless*' means 'having no past end-point', etc. (In the context of Penrose (1968) and Hawking & Penrose (1970), 'past-endless' is equivalent to 'past-inextendible'.) Note that the existence of a timelike curve with end-points u and v is both necessary and (less trivially) sufficient for the existence of a timelike curve containing u and v.

[4] This concept is actually self-dual, for the boundary of a past-set P is simultaneously the boundary of the future-set $I^+[M-P]$. Such sets are sometimes called 'achronal' or 'semi-spacelike boundaries'.

[5] A set is *achronal* if no two of its points can be joined by a timelike curve.

[6] The *future Cauchy horizon* $H^+(S)$ of an achronal set S is the set $D^+(S) - I^-[D^+(S)]$. The boundary in the manifold $I^+[S]$ of the set $D^+(S) - S$ is then $H^+(S) - S$; and it follows from theorems (2.5) and (2.8) that, in this manifold, $H^+(S) - S$ is ruled by past-endless null geodesics.

[7] Observe that (2.2) yields an immediate proof that past-distinguishing space-times contain no self-intersecting causal curves.

[8] Take the set $\Omega \times (0, 1]$, ordered lexicographically, where Ω denotes a well-ordered uncountable set each element of which has only a countable set of predecessors, and where the half-open interval $(0, 1]$ is given its usual ordering.

[9] ***FC 3*** states that X is a *past-reflecting* $\mathfrak{B}$*-space* (Kronheimer & Penrose 1967). The converse of ***FC 3*** holds for all causal spaces.

[10] For some space–times (e.g. Minkowski space) the space itself is both an IP and an IF, so $\hat{M}$ and $\check{M}$ are not disjoint. In such cases we should, to be accurate, here take a *disjoint* union of $\hat{M}$ and $\check{M}$. To avoid the consequent (purely notational) complication, we shall, however, phrase our argument as though no subset of M were simultaneously both an IP and an IF.

REFERENCES

Carter, B. 1966 *Phys. Rev.* **141**, 1242.

Courant, R. & Hilbert, D. 1965 *Methods of mathematical physics: II.* New York: Interscience.

Geroch, R. 1968 *J. Math. Phys.* **9**, 450.

Geroch, R. 1970*a* *J. Math. Phys.* **11**, 437.

Geroch, R. 1970*b* In *Relativity* (ed. M. Carmeli, S. Fickler and L. Witten). New York: Plenum Press.

Graves, J. C. & Brill, D. R. 1960 *Phys. Rev.* **120**, 1507.

Hawking, S. W. 1966*a* *Proc. R. Soc. Lond.* A **294**, 511.

Hawking, S. W. 1966*b* *Proc. R. Soc. Lond.* A **295**, 490.

Hawking, S. W. 1967*a* *Proc. R. Soc. Lond.* A **300**, 187.

Hawking, S. W. 1967*b* *Singularities and the geometry of space–time* (Adams Prize Essay, Cambridge University).

Hawking, S. W. & Penrose, R. 1970 *Proc. R. Soc. Lond.* A **314**, 529.

Kronheimer, E. H. 1972 (to appear).

Kronheimer, E. H. & Penrose, R. 1967 *Proc. Camb. Phil. Soc.* **63**, 481.

Kruskal, M. D. 1960 *Phys. Rev.* **119**, 1743.

Penrose, R. 1964 In *Relativity, groups, and topology* (ed. B. and C. DeWitt). New York: Gordon Breach.

Penrose, R. 1965*a* *Phys. Rev. Lett.* **14**, 57.

Penrose, R. 1965*b* *Proc. R. Soc. Lond.* A **284**, 159.

Penrose, R. 1968 In *Battelle rencontres* (ed. C. DeWitt and J. A. Wheeler). New York: Benjamin.

Penrose, R. 1972 In *Techniques of differential topology in relativity.* Philadelphia: SIAM.

Rindler, W. 1956 *Mon. Not. R. astr. Soc. Lond.* **116**, 662.

Schmidt, B. 1971 *Gen. Relativity Gravitation* **1**, 269.

Seifert, H.-J. 1968 Kausale Lorentzräume (Thesis, Hamburg University).

Seifert, H.-J. 1971 *Gen. Relativity Gravitation* **1**, 247.

Synge, J. L. 1950 *Proc. R. Irish Acad.* A **53**, 83.

51

BLACK HOLES

Scientific American **226(5)** (1972), 38–46

BLACK HOLES

If a massive star collapsed to a sufficiently small volume, light could not escape from it. A rotating black hole could account for the radiation of gravitational waves from the center of the galaxy

by Roger Penrose

In about five billion years the sun will have consumed so much of its hydrogen in thermonuclear reactions that it will evolve to a star of the type known as a red giant. Standard theory predicts that the sun will grow to some 250 times its present diameter of 850,000 miles, devouring Mercury, Venus and probably the earth in the process. Its substance will then be only a tenth as dense as air. (At present the average density of the sun is a fifth the density of the earth.)

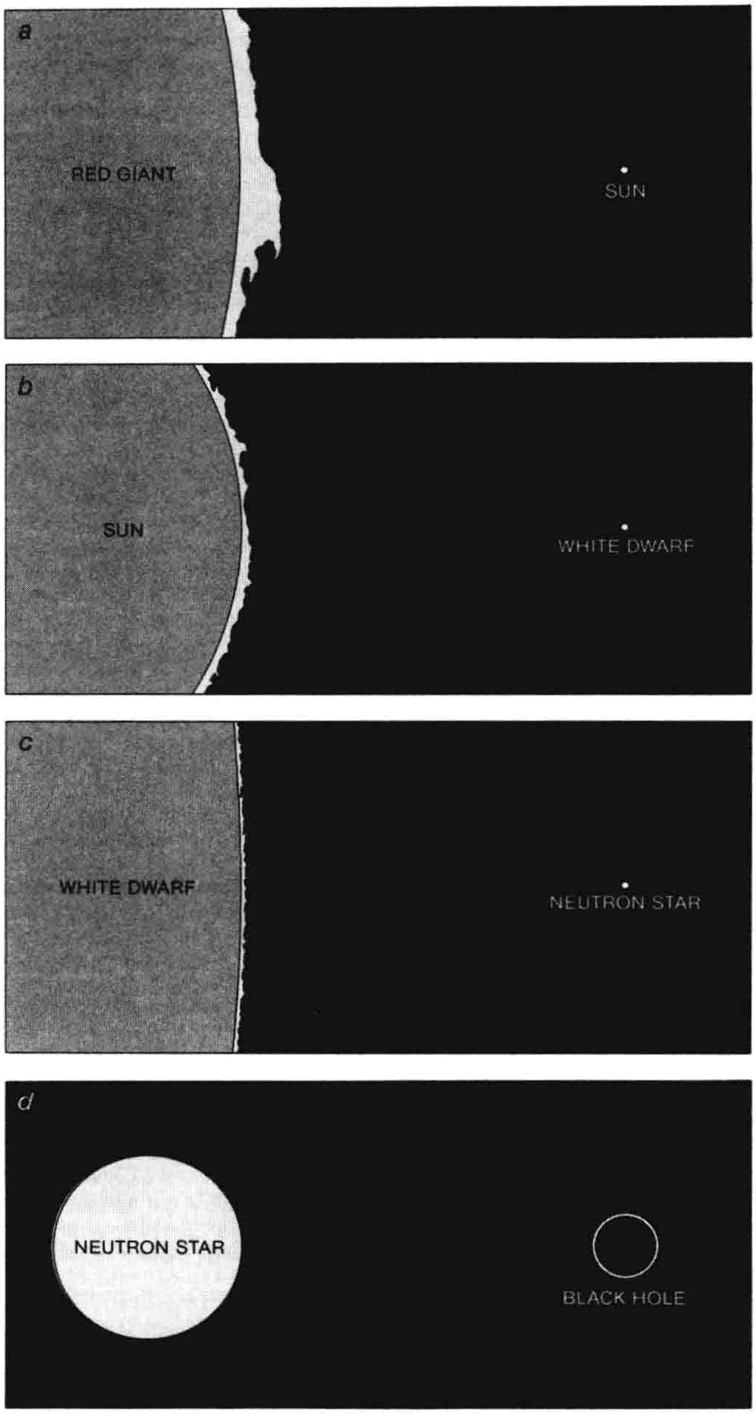
As it consumes more and more of the available nuclear fuel (helium and heavier elements as well as hydrogen), the bloated sun will reverse its expansion and contract down through its present diameter to a mere one-hundredth of that diameter, or roughly the size of the earth. It will then have evolved into a white dwarf and its contraction will halt. The electrons of its atoms will have become packed together so tightly that one of the laws of quantum mechanics will come into play and give rise to an effective pressure strong enough to prevent further contraction. The law is the Pauli exclusion principle, which states that no two electrons can occupy the same energy state. By this time the sun's density will be enormous; a ping-pong ball filled with its substance would have the mass of several elephants. All that remains for it is to cool off to its final dead state as a black dwarf.

No substance on the earth has a density remotely approaching the density of a white dwarf. Nevertheless, many white dwarfs (and red giants) are observed in the heavens. They are part of the evolutionary history of perfectly normal stars such as the sun. Moreover, the theory of stellar evolution to a white dwarf fits these observations closely. Not all stars, however, can follow this "normal" evolutionary path. In Subrahmanyan Chandrasekhar's theoretical investigations of stellar structure in 1931, he found that there was a maximum mass above which a white dwarf cannot sustain itself against further gravitational contraction. The gravitational pull toward the star's center would be even greater than the counteracting pressure from the action of the Pauli exclusion principle on the electrons. This maximum-mass limit is not much greater than the mass of the sun. Chandrasekhar's original limit was about 1.4 solar masses; more recent computations give an even lower value. Many observed stars have a mass greater than 1.5 solar masses. What can be their ultimate fate?

Consider a star that has twice the mass of the sun. Like the sun, it will expand to enormous size after consuming much of its original hydrogen fuel. Then it will contract again. It has no stable equilibrium state that will enable it to settle down as a white dwarf. Therefore the star, or a substantial portion of it, will collapse through the white-dwarf dimension. Owing to the extreme temperatures and densities that are involved, it will undergo processes that lead it to explode catastrophically. Such exploding stars, called supernovas, have been observed in our galaxy (the last one was described by Johannes Kepler in 1604) and in other galaxies. A supernova will outshine even an entire galaxy for a few days. Perhaps as much as 90 percent of the star's mass could be ejected. Only a collapsed core of the star would remain at the center of a rapidly expanding cloud of gas. (The Crab Nebula is such a cloud.) The core is much too small and compressed to form a white dwarf; it can find equilibrium only as a neutron star.

Even by comparison with a white dwarf a neutron star is tiny. The reduction from a white dwarf to a neutron star is even more than the reduction of 100 to one from the sun to a white dwarf, or probably rather more than the approximate reduction of 250 to one from a red giant to the sun. A neutron star may be only 10 kilometers in radius, or only about one seven-hundredth the radius of a white dwarf. The density of a neutron star could be more than 100 million times the extraordinary density of a white dwarf. A ping-pong ball filled with material from a neutron star would have the mass of a large asteroid such as Juno, a minor planet 118 miles across. The star's density would be comparable to the density of a proton or a neutron; indeed, a neutron star can be regarded as an oversized atomic nucleus, the only essential difference being that it is bound together by gravitation rather than by nuclear forces. Most of its electrons have been pushed into its protons, with the result that the protons have become neutrons. It is now the Pauli exclusion principle acting on the neutrons that supplies the effective force preventing further collapse.

This picture of a neutron star was predicted theoretically by J. Robert Oppenheimer, Robert Serber and G. M. Volkoff in 1938 and 1939. For many years astronomers doubted that neutron stars could actually exist. Since 1967 the observational situation has changed drastically. In that year the first pulsars were observed. The theory of pulsars has developed rapidly since then, and it now seems virtually certain that the radio and optical pulses emitted by pulsars owe their energy and extraordinary regularity to the presence of a rotating neutron star. At least two pulsars reside inside supernova remnants, one being the Crab Nebula, giving further support to the theory that pulsars are in fact neutron stars.



COMPARATIVE SIZES of a red-giant star, the sun, a white-dwarf star, a neutron star and a black hole are illustrated. The red giant (*a*) has approximately the same mass as the sun, yet its diameter of about 200 million miles is 250 times larger than the sun's. The sun (*b*) has 100 times the diameter of a white dwarf of the same mass. The diameter of a white dwarf (*c*), which is roughly the same as the diameter of the earth, is about 700 times greater than the diameter of a neutron star with the mass of the sun. The neutron star has only to collapse to a third of its diameter to form a black hole (*d*). The black hole is no more than four miles across although it has the mass of the sun. Its radius is proportional to its mass.

As in the case of white dwarfs, there is a maximum possible mass above which a neutron star would not be able to sustain itself against still further gravitational contraction. There is some uncertainty as to the exact value of this maximum-mass limit. The original value given by Oppenheimer and Volkoff in 1939 was about .7 solar mass. More recently values of up to three solar masses have been suggested. These higher values take into account that the heavy subatomic particles called hyperons may be present in addition to ordinary neutrons and protons. In any case the correct limit could not be more than a few solar masses, and there are stars whose mass is more than 50 times the mass of the sun. What can be their eventual fate? It is very unlikely that such stars would, either in a final collapse or at some earlier stage, inevitably throw off so much of their material that their mass would always fall below the limit required for a stable white dwarf or a neutron star. What other forms of condensed matter could exist, with densities even in excess of the fantastic value that is maintained inside a neutron star?

Theory tells us that although greater densities can be achieved, it is not possible to obtain any further stable equilibrium state. The gravitational effects become so overwhelming that they dominate everything else. Newtonian gravitational theory becomes quite inadequate to handle the problem and we must turn to Einstein's general theory of relativity. In doing so we are led to a picture so strange that even a neutron star seems commonplace by comparison. This new picture, originally put forward by Oppenheimer and Hartland S. Snyder, has earned the appellation "black hole."

A black hole is a region of space into which a star (or a collection of stars or other bodies) has fallen and from which no light, matter or signal of any kind can escape. What further contraction is necessary for a neutron star to be compressed to the size of a black hole? Consider an object of one solar mass. We have seen that the sun has a diameter some 70,000 times greater than the diameter of a neutron star, and that a red giant has a diameter about 20 million times greater than the diameter of a neutron star. In view of these large differences in size it is perhaps surprising to find that a neutron star would have to contract only to approximately a third of its diameter to become a black hole. Larger black holes are also possible, but they would be the result of the collapse of a star, or collection of bodies, whose final total mass was greater than the sun's. The linear dimension of the hole is proportional to the hole's mass.

The general theory of relativity plays a significant role in the theory of neutron stars, even before the extreme conditions of a black hole are reached. Physical theory has worked well in the description of stars of enormously different sizes and densities. From this point of view there would seem to be little reason to doubt the slight extrapolation needed to cover the case of a black hole. Such a view, however, is not quite fair. The part of physical theory on which the detailed description of a black hole depends, namely the general theory of relativity, cannot be said to have played an irreplaceable role in observational astronomy. The possibility that the general theory of relativity is wrong must be seriously considered. The observational and experimental tests of the general theory that have been successfully conducted are still not very numerous. Although there is no conflict between the theory and the observations, these observations still do not point conclusively in the direction of the general theory. There is still scope for alternative theories of gravitation.

It must nonetheless be said that the general theory of relativity is an excellent theory; it is almost certainly the most satisfactory theory of gravity available to us. Furthermore, the Brans-Dicke-Jordan scalar-tensor theory, which can be regarded as the most serious rival to the general theory, leads to a black-hole picture identical with the one arising from Einstein's theory. Even in Newtonian theory a situation similar to a black hole can arise. As early as 1798 Pierre Simon de Laplace predicted from Newtonian mechanics that a sufficiently massive and concentrated body should be invisible because the escape velocity at its surface would be greater than the speed of light. Therefore a photon, or particle of light, emitted radially from the surface would fall back to the surface and so not escape to be observed at large distances from the body. This description is perhaps arguable, but it shows that there is a situation to be faced even in Newtonian theory. Having made these observations, however, I shall now restrict the discussion to considerations totally within the bounds of the general theory of relativity.

To begin with, let us consider the present standard picture of a black hole. The black hole is characterized by a spherical surface whose radius is proportional to the hole's mass. This surface is called the "absolute event horizon"; its defining property is that signals emitted inside it cannot escape, whereas from any point outside it signals can be emitted that do escape. The size of the sphere for any mass, that is, the radius of the event horizon, can be calculated by multiplying twice the mass by the universal gravitational constant and dividing the result by the speed of light squared ($2mG/c^2$). Performing this calculation for the sun yields the result that the sun would have to be collapsed into a sphere four miles in diameter; the absolute event horizon would be the surface of this four-mile sphere.

The body whose collapse was responsible for the black hole's existence has fallen deep inside the event horizon. The gravitational field inside the event horizon has become so powerful that even light itself is inevitably dragged inward regardless of the direction in which it is emitted. Outside the event horizon light escapes if it is aimed suitably outward. The closer the emission point is to the event horizon, the more the wave front of the emitted signal is displaced back toward the center of the black hole. We may intuitively regard this displacement as being caused by the effect of the gravitational attraction on the motion of the light. The light seems to travel more easily in the direction of the gravitating center of the black hole than it does in the outward direction. Inside the event horizon the inward pull has become so strong that the outward motion becomes altogether impossible. On the event horizon itself it is possible for light to "mark time," forever hovering at the same distance from the center of the black hole.

This behavior applies not only to light but also to any signal or material body. The velocity of light is still the limiting velocity even inside the event horizon. The special theory of relativity still holds locally, although it is not evident from this picture. The local frames of reference, in terms of which the special theory of relativity is normally described, are themselves falling rapidly inward to the gravitational center.

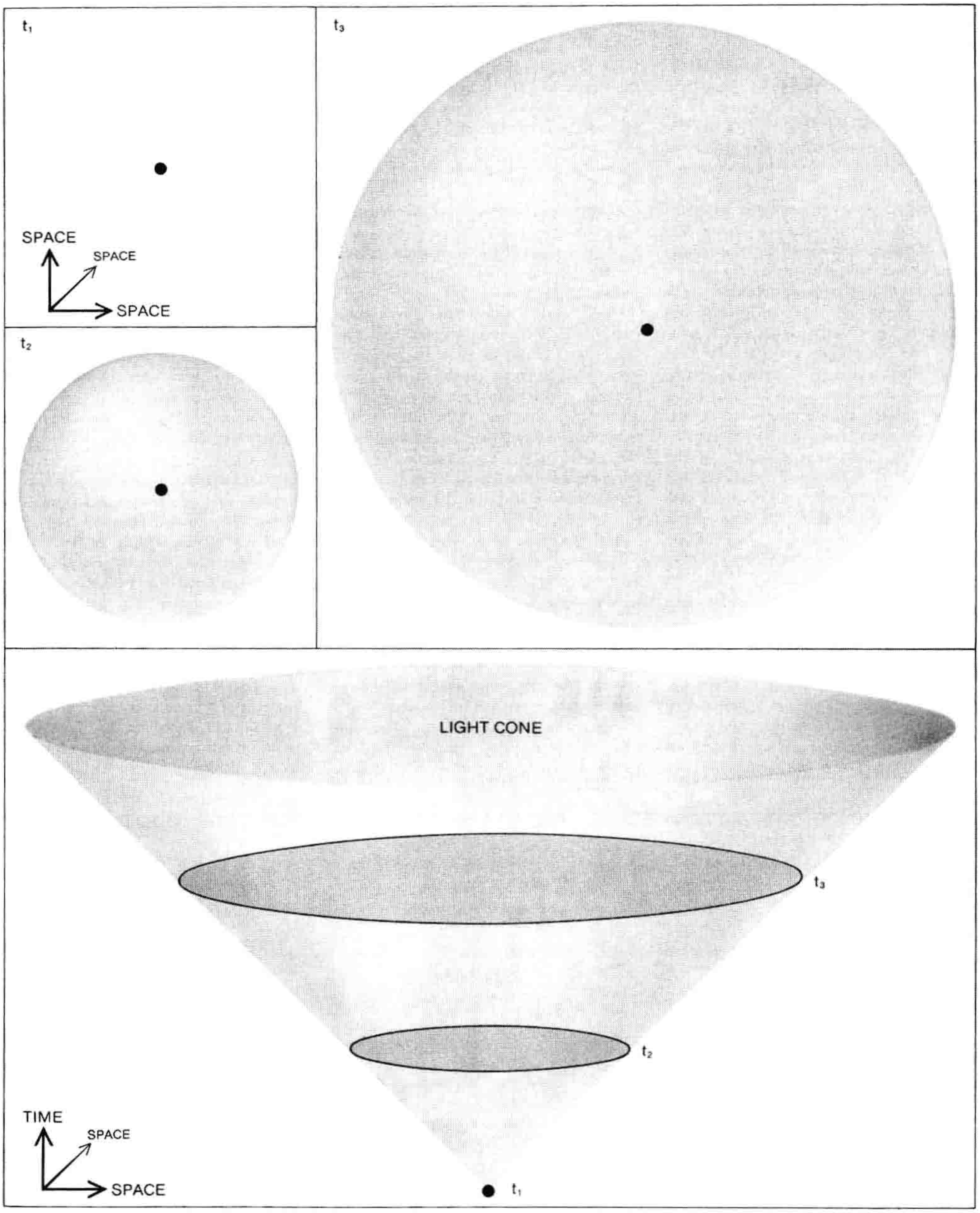
It is much more satisfactory to give a space-time description of the situation rather than the purely spatial one given so far. A space-time representation suppresses one of the spatial coordinates and substitutes a time coordinate. It gives an instantaneous picture of what is going on at all times, obviating the need for many sequential "snapshots" of a developing situation.

Consider a light flash emitted in all directions from a given point in ordinary space. The wave front of the flash would be a sphere centered on the emitting point and growing larger each moment at the speed of light. A purely spatial representation of the flash would be a sequence of spheres, each sphere larger than the preceding one, marking the position of the light flash's spherical wave front at a given moment in time. A space-time representation of the light flash, however, would be a cone whose vertex represents the time and place at which the light flash is emitted, the cone itself describing the history of the light flash.

By the same token the history of a star's collapse down to a black hole can best be depicted in a space-time representation [*see illustration on page 43*]. The locations of the light cones at various points in space-time show how light signals propagate in the gravitational field. At some points the light cones are drawn as being tipped over, but this is not something that would be noticed by a local observer. Such an observer would follow a path in space-time that proceeds into the interior of the light cone; his speed can never be greater than the speed of light, and only inside the light cone is this criterion met. By taking a horizontal section through the space-time picture we get a spatial representation of the situation [*see illustration on page 42*].

What is the fate of the original body that collapsed to produce the black hole? Assuming that exact spherical symmetry is maintained right down to the center, the answer provided by the general theory of relativity is a dramatic one. According to the general theory, the curvature of space-time increases without limit as the center is approached. Not only is the substance of the original body squeezed to infinite density at the center of the black hole—that is, effectively crushed out of existence—but also the vacuum of space-time outside the body becomes infinitely curved. The effect of this infinite curvature on a hapless observer, were he foolhardy enough to follow the body inward, would be catastrophic. He would feel tidal forces across his body that would mount rapidly and would reach infinity within a finite period of his experienced time.

The gravitational tidal effect is the most direct physical manifestation of space-time curvature. Einstein pointed out that the gravitational force on a body can be eliminated at any one point simply by choosing a frame of reference that is falling freely. He gave the famous



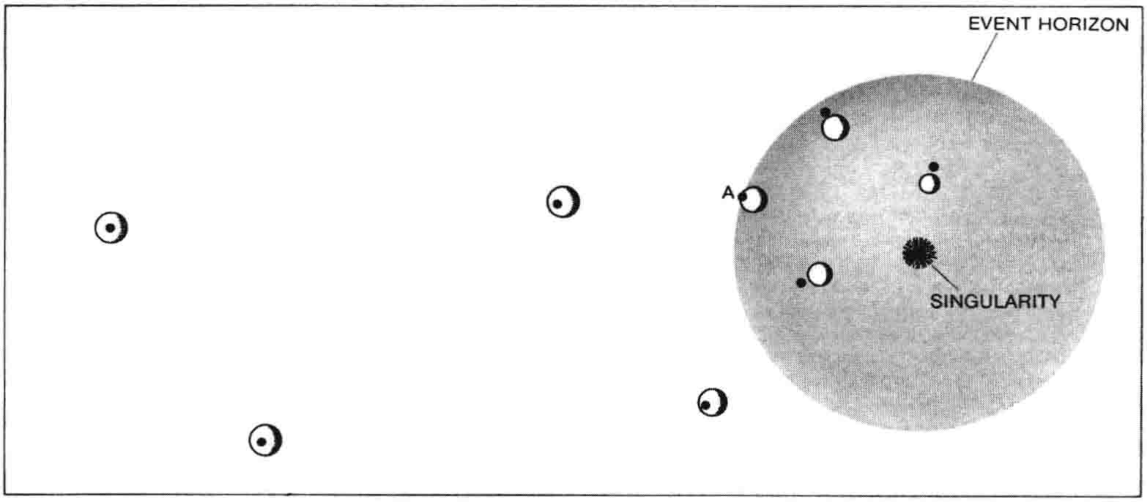
PURELY SPATIAL REPRESENTATION of the history of a light flash emitted from a point would be successive snapshots of the light's spherical wave front expanding through space (*top*) at given instants of time: t_1, t_2, t_3 and so forth. A space-time representation of the same event would be a light cone with its apex at the point emitting the light (*bottom*). In order to picture a light cone, one of the spatial dimensions must be suppressed. The space axes are depicted in the horizontal plane; the time axis runs vertically with time increasing upward. The circles (*color*), foreshortened to ellipses, intersecting the light cone are the instants t_1, t_2, t_3 at which the snapshots were taken in the spatial representation. The space-time light cone shows in one diagram the continuous history of the expanding wave front of the light signal, obviating the need for many sequential snapshots of the developing situation.

example of an elevator that broke its cable and fell toward the earth. Any passenger inside would be falling at the same rate as the elevator. He would therefore feel no net gravitational acceleration and would float free of gravity inside the elevator. Such elimination of the gravitational force by free fall is now a familiar feature of space travel. The tidal effect, however, cannot be so eliminated and is therefore a true manifestation of the gravitational field. Imagine an observer falling freely in the earth's field. Suppose he is surrounded by a sphere of particles that he initially observed to be at rest with respect to himself. The Newtonian gravitational field of the earth varies as the inverse square of the distance between it and any other body, pulling more strongly on objects closer to its surface than on objects farther away. This nonuniformity of the gravitational field tidally distorts the sphere of particles into an ellipsoid [*see top illustration on page 44*]. The earth's ocean tides are an example of the effect; in this case it is the earth experiencing the tidal force of the moon.

The tidal effects encountered in the solar system are small; the largest actually occurs at the surface of the earth itself and is due primarily to the earth's gravitational field. These tidal effects are not noticeable at all on a laboratory scale. To put it another way, the curvature of four-dimensional space-time at the earth's surface is not significant over laboratory dimensions. The degree of space-time curvature can be stated in terms of radius of curvature. The smaller the curvature of space-time, the larger the corresponding radius of curvature, just as in three dimensions the smaller the degree to which the surface of a sphere is curved, the larger its radius of curvature. The radius of space-time curvature at the surface of the earth is approximately the distance of the earth from the sun. (That is a pure coincidence; the sun is itself irrelevant to this particular tidal effect.) Thus the earth does not curve space-time much compared with laboratory dimensions. The tidal effect at the surface of the sun is even less because of the sun's lower average density. In fact, the radius of space-time curvature at the sun's surface is approximately twice the distance from the earth to the sun. Hence the sun curves space-time less at its surface than the earth does at the earth's surface.

At the surface of a white dwarf the degree of space-time curvature is much greater; the radius of curvature is of the same order as the radius of the sun: about 500,000 miles. The tidal effects near a white dwarf would be very noticeable to an astronaut in orbit around the star. The man's head and feet would experience forces in different directions of perhaps a fifth of the total force he would feel standing on the earth. At the surface of a neutron star, however, the tidal effects are by ordinary standards enormous. There the radius of space-time curvature is only about 30 miles. It is clear that no astronaut in a low orbit around a neutron star could possibly survive; even if he curled himself into a small ball, the gravitational acceleration at various parts of his body would differ by several million times from the gravitational acceleration (g) at the surface of the earth.

In principle instruments could be built to withstand such tidal forces. They would have to be built very small to keep the forces down, since the tidal force is proportional to the size of the object on which it acts. Imagine now that such an instrument fell into a black hole of one solar mass. As it crossed the event horizon it would experience tidal forces about 30 times the tidal forces it would feel at the surface of a neutron star. The instrument could remain intact, however, since the forces on its constituent parts would still be small. The tidal forces would rise rapidly as the instrument approached the center of the black hole, ripping apart in turn the material of the instrument, the molecules of which the material was composed, the atoms that constituted the molecules, the nuclei of the atoms and finally even the elementary particles that a moment previously had been the building blocks of the nuclei. Moreover, the entire process would not last more than a few milliseconds. What we have is a model of the



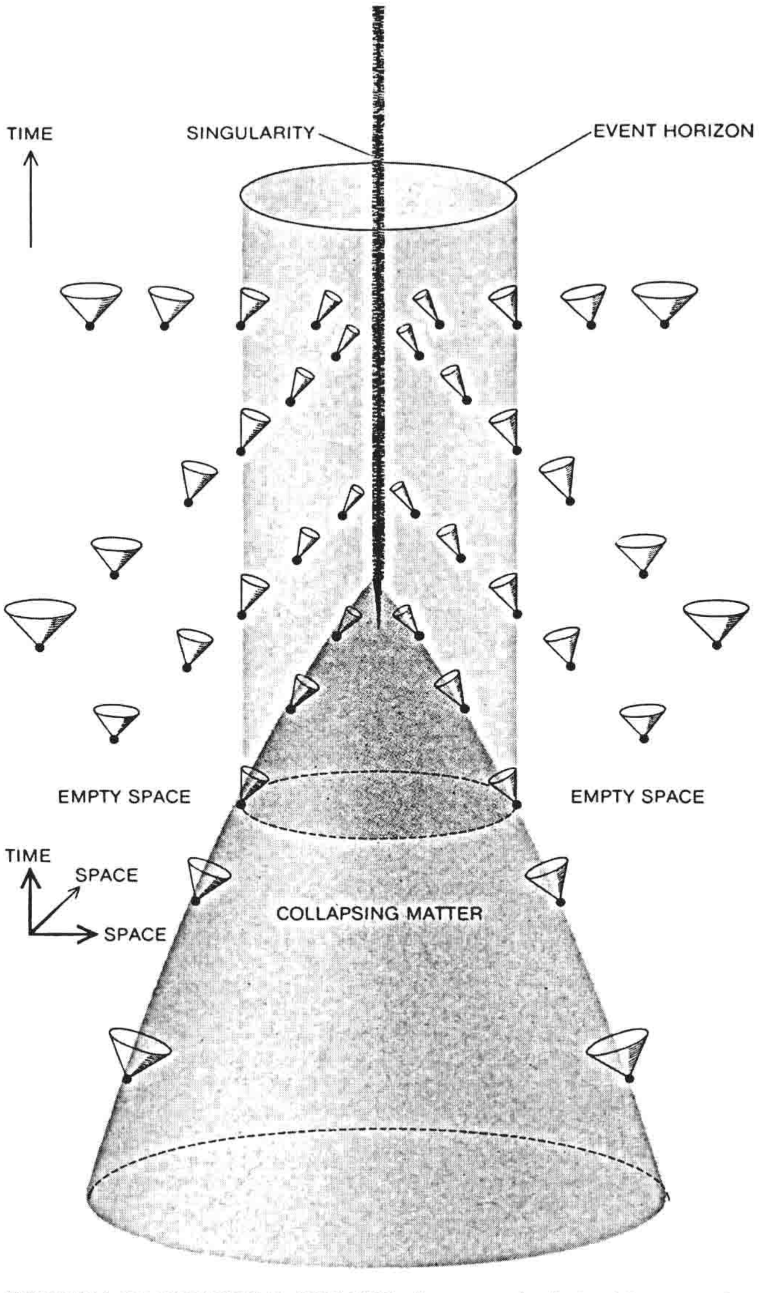
EVENT HORIZON (*gray sphere*) is the boundary of the black hole. Signals cannot escape from the inside of the event horizon but must propagate toward its center. In this spatial representation a light signal originating at any one of the points shown occupies at the next instant a spherical surface (*small circles*) known as its wave front. At large distances from the black hole the point lies at the center of this surface. At smaller distances the spherical surface is displaced by the strong gravitational attraction of the black hole. The spherical wave front of a point exactly on the event horizon touches the event horizon internally and never escapes (*A*).

creation of the universe in reverse and on a much smaller scale. The initial "big bang" of the cosmological models is what is called a space-time singularity: the curvature of space-time becomes infinite. The situation inside a black hole leads also to such a singularity, but of a time-reversed character.

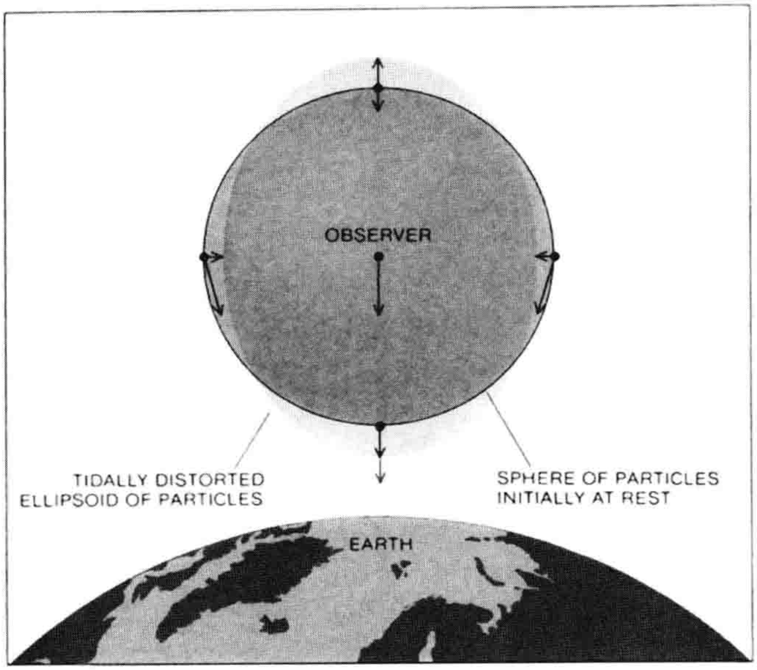
There are many questions that must be raised concerning the acceptability of this picture as a realistic description of something that actually happens in nature, quite apart from any doubts one might have about the validity of the general theory of relativity. In the first place, can we be sure that enough is known about the nature of matter under the extreme compressions required to form a black hole for these predictions to carry conviction? What role does the assumption of exact spherical symmetry play in the discussion? To what extent does the black-hole picture fit in with astronomical observations? Let us consider these questions in turn.

The question concerning the excessive densities—densities somewhat greater than that of nuclear matter—involved in the creation of a black hole is not nearly as serious as it might seem at first. Even if it were argued that present physical knowledge is inadequate at such densities (which is probably not the case), the argument could only shed doubt on the collapse picture for the least massive of collapsing bodies. The density of any body is proportional to its mass divided by its radius cubed; the radius of a black hole is proportional to its mass. These two facts mean that the density at which a black hole is formed would be inversely proportional to the square of the mass.

A black hole with a mass ranging from 10,000 to 100 million solar masses is considered a possible candidate for what might inhabit the center of a galaxy. A collapsing mass equal to 100 million solar masses would reach a black-hole situation when its average density was roughly the density of water. The tidal effects at the event horizon are also proportional to the inverse square of the mass of the hole. Thus for a body of 100 million solar masses these tidal effects would be somewhat less than those produced at the earth's surface. An astronaut could pass through the event horizon without the tidal forces' affecting him. He would not notice anything particular happening as he crossed the event horizon. (In any case the exact location of the horizon is not something that can be discerned by local measure-



SPHERICAL GRAVITATIONAL COLLAPSE of a star can be depicted in a space-time diagram in which two of the three space dimensions are plotted horizontally and the time dimension is plotted upward. The star collapses until its gravitational field becomes so strong that light cannot escape, creating an event horizon. The matter of the star collapses down to a "singularity" of zero volume and infinite density where the present laws of physics break down. The propagation of light signals from various points is indicated by light cones. The light from points closer to the strongly attracting singularity is displaced more toward the singularity than light from points farther away. Even though the light cones are depicted tilted near the singularity, the special theory of relativity still holds for the local frames of reference of the light cones and the speed of light is still the limiting velocity. Compare this space-time representation of the behavior of light near the event horizon with the purely spatial representation on the opposite page. The spatial representation may be taken to be a horizontal section through the above space-time picture.


OBSERVER
TIDALLY DISTORTED ELLIPSOID OF PARTICLES
SPHERE OF PARTICLES INITIALLY AT REST
EARTH

GRAVITATIONAL TIDAL EFFECT is the most direct manifestation of space-time curvature. If an observer falls freely in the earth's gravitational field, he feels no net force. If he is surrounded by a sphere of particles (*color*) that are initially at rest in relation to him, he can notice the nonuniformity of the Newtonian gravitational field as a tidal effect that distorts the sphere into an ellipsoid (*gray*). Tidal effect is a true manifestation of the gravitational field and cannot be eliminated by a change in the observer's frame of reference.

ments.) And the astronaut would have a few minutes to enjoy the experience of life inside a black hole before the tidal effects mounted to infinity. For a black hole of 10 billion solar masses he would have about a day. As for the large tidal effects and densities encountered inside the black hole, this is a matter that involves certain implications of the general theory of relativity, to which I shall return.

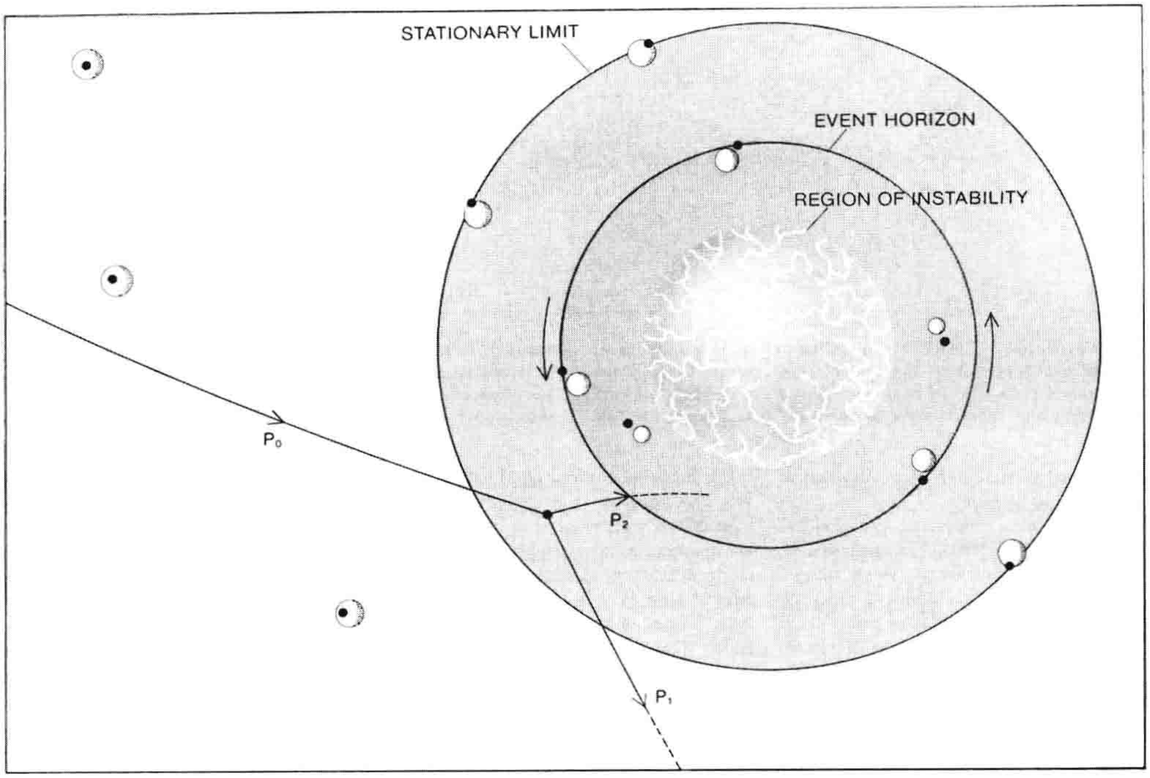
The question concerning the role played by the assumption of spherical symmetry is much more serious. If we do not assume spherical symmetry, then we cannot appeal to the exact solutions on which we have based the foregoing discussion. Furthermore, even if we assume that initially the deviations from spherical symmetry are slight, we should have every reason to expect that near the central point these asymmetries would be enormously magnified. Might not the different portions of the collapsing body miss one another? Perhaps they could reemerge after a close encounter and bounce out again. Even if they did not, can we say anything about the final configuration of the gravitational field resulting from the collapse? It is fortunate that, owing to some general theorems that have been proved over the past several years, a complete picture of asymmetrical collapse has emerged.

Considering the picture in a little detail, suppose that a massive star or a collection of bodies collapses and that deviations from spherical symmetry are at first comparatively small. We can establish that a point of no return has been passed if a certain criterion is satisfied. That criterion can be stated in several different ways, but the following is the simplest. Imagine that a flash of light is emitted at a point in space-time. If the event is depicted in a space-time representation, the flash of light will be the light cone of the point. The light rays start out from the point by diverging in all directions. When they pass through matter or through a gravitational field, the matter or the field has a focusing effect on the rays. If enough matter or a sufficiently strong gravitational field is encountered, the amount that the rays diverge can be reduced to such an extent that it is actually reversed, that is, the rays start to converge. The required criterion for a point of no return is that every light ray from the space-time point encounter enough matter or gravitation for the light cone to be reconverged. It is not hard to show from simple order-of-magnitude estimates that, for sufficiently large collections of mass, the criterion can indeed be satisfied before densities or curvatures become excessive, and without any assumption of symmetry.

Once this criterion is satisfied, a number of things can be said. In the first place, according to a precise theorem in the general theory of relativity put forward by Stephen Hawking and the author, it follows that somewhere there must be a space-time singularity. A singularity is a region of space-time where physical theory breaks down. Here the singularity is a region where we must assume that infinitely strong gravitational tidal forces deform and squeeze matter and photons out of existence. Physicists are unhappy with a theory that predicts the evolution of a truly singular physical state for matter. In the past whenever a singularity was encountered in a theory, it was generally a warning that the theory in its present form was breaking down and new theoretical tools were needed. In the case of black holes we are again being presented with a situation of this kind, but one more serious than before. Here the singularity refers to the very structure of space and time.

There are two distinct possibilities at this stage. It may be that the resulting singularity is such that signals can escape from it and be observed at large distances. This is the more alarming of the two possibilities, and it is also the more conjectural. Such a singularity is called naked. The possibility of naked singularities is alarming because the physical effects of near-infinite space-time curvatures are quite unknown. If these effects can influence the outside world, then an essential uncertainty is introduced into physical theory.

On the other hand, it is possible that the singularities resulting from gravitational collapse are always hidden from view, as was true in the spherically symmetrical situation. In this case the essential uncertainty does not arise. The assumption that only this less alarming possibility can arise in gravitational collapse is what is called the hypothesis of "cosmic censorship." This hypothesis simply forbids a naked singularity. There is perhaps some slight theoretical evidence in favor of the hypothesis, but

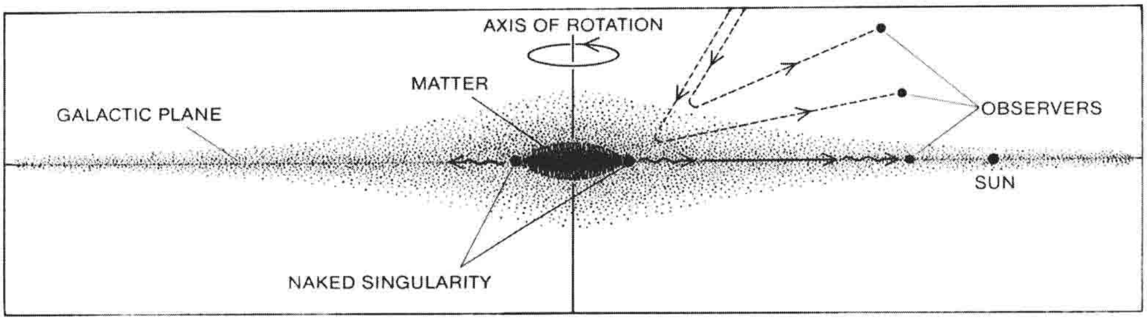


ENERGY CAN BE EXTRACTED from a rotating black hole even though no material object can escape once it has fallen below the event horizon (*solid gray*). In the case of a rotating black hole the event horizon rotates. The stationary limit is a spheroidal surface (*color*) on which it is necessary for a signal or a body to travel with the speed of light in order to "stay in the same place" as viewed from infinity. If a particle, P_0, drops from infinity to within this surface, it is possible for it to divide in two in such a way that one component, P_2, falls into the black hole but the other component, P_1, escapes back to infinity with more mass-energy than the initial particle P_0. In this way P_1 extracts some of the rotational energy of the black hole. The location of the internal singularity is uncertain when perturbations in addition to rotation are present, but it is most likely to be along or inside central region of instability.

the situation is inconclusive at the moment. I would personally tend to believe in the hypothesis in situations that initially do not differ too much from spherical symmetry. In more extreme situations the question is quite open. Perhaps there is even observational evidence against the hypothesis of cosmic censorship.

If we assume that the hypothesis is indeed true, then much can be said. Once the focusing criterion is satisfied, the hypothesis of cosmic censorship implies that an absolute event horizon will arise. This horizon will have a well-defined area of cross section that has a tendency to increase with time and can never decrease. Thus black holes tend to grow and cannot shrink. It seems reasonable to suppose a black hole, if left undisturbed, would settle into a stationary state. One might think that such stationary states of black holes could be very complicated owing to the large number of possible configurations of matter that could collapse down to the black-hole size. Some remarkable recent work by Werner Israel, Brandon Carter and Hawking has shown that this is not the case. Only a very restricted class of stationary black-hole configurations can arise. They are uniquely characterized by the value of the mass, spin and charge of the hole. Einstein's equations for the general theory of relativity have been solved explicitly by Roy P. Kerr and Ezra Newman for these configurations. The reason the asymmetries present in the collapsing body do not show up in the final state of the black hole is that once the hole is formed the body that produced it has little influence on its own subsequent behavior. The black hole is best thought of as a self-sustaining gravitational field governed by the internal dynamics of the general theory of relativity. These dynamics allow the asymmetries in the gravitational field of the hole to be carried away in the form of gravitational waves.

We have seen that a material object, once swallowed by a black hole, cannot escape. On the other hand, there are mechanisms whereby some of the energy content of the black hole can be extracted. One such mechanism would allow two black holes to coalesce. This process would be accompanied by the copious emission of gravitational waves, whose total energy should be a substantial fraction of the initial rest-mass energy of the black holes. Another mechanism would be to allow a particle to fall into a region close to the event horizon of a rotating black hole. The particle splits into two particles in such a way that one falls into the hole and the other escapes back to infinity with more mass-energy than the initial particle had. In this way rotational energy of the black hole is transferred to the particle mo-



ROTATING SINGULARITY may explain the copious emission of gravitational waves from the center of the galaxy. Calculations show that certain types of naked singularity can be observed only from one plane. Lines of sight "bounce out" if they are not in this plane. If such a singularity were the source of gravitational waves, it would beam them in one plane. If this plane happened to be the plane of the galaxy, then the sun and the earth, which are also in the galactic plane, would be in the path of emission of the waves.

tions outside the hole [*see illustration on preceding page*].

Let us now consider the situation inside the black hole and the existence of the space-time singularity. Since "singularity" means a region where physical theory breaks down, we have the curious situation here that the general theory of relativity is predicting its own downfall. Perhaps we should not be too surprised, however, since we are treating the general theory of relativity only in its capacity as a classical theory. It is to be expected that when space-time curvatures become sufficiently enormous, quantum effects must play a dominant role. When the radius of space-time curvature becomes as small as, say, 10^{-13} centimeter (roughly the radius of an elementary particle), then the theory of particle physics, as it is currently understood, must break down. If the radius of space-time curvature ever becomes as small as 10^{-33} centimeter, we must apply quantum mechanics to the structure of space-time itself. At present there is no satisfactory theory for such application.

The final question is: What is the observational status of black holes? A number of claims and counterclaims have been made over the past year or so, but at present it seems that no unambiguous conclusions can be drawn. The main discussion centers around double-star systems in which one component is the suspected black hole and around multiple-star systems such as a globular cluster. The presence of a black hole could be ascertained by its gravitational influence on other bodies, affecting their motions. If an invisible object is thereby detected whose mass is too large for it to be a white (or black) dwarf or neutron star, then the case for its being a black hole is a strong one.

There is another aspect to the role of black holes in observational astronomy. We can compare the situation to the case of neutron stars. For many years astronomers had attempted to detect neutron stars by searching for certain effects, such as X-ray emission, that had been predicted to accompany their presence. When neutron stars were finally detected, they were discovered by an effect that was totally unexpected and is still not really satisfactorily explained: the emission of rapid, regular, sharp pulses of electromagnetic radiation characteristic of a pulsar. It is quite possible that the detection of black holes will also come about because of some equally unexpected side effect. There is no shortage of unexplained phenomena in astronomy today that might conceivably be relevant. There is the apparently phenomenal energy output of quasars and radio galaxies, the explosions in the center of galaxies, the anomalous red shifts in the spectra of some quasars and galaxies, and the discrepancies in the measurements of the mass of galaxies. Serious questions remain even about the spiral-arm structure of normal galaxies. Above all, there is the apparent observation by Joseph Weber of the University of Maryland of gravitational waves emanating from the center of our galaxy [see "The Detection of Gravitational Waves," by Joseph Weber; SCIENTIFIC AMERICAN, May, 1971]. If these waves are emitted continuously in all directions from the galactic center, the energy they carry would result in a loss of mass from the galaxy of many thousands of solar masses per year. Such a number would seem to be in gross contradiction to other observations.

The theory of black holes has not led to a convincing explanation of this phenomenon or of any of the other phenomena mentioned above, but the subject is still young. In the case of Weber's waves the best hope would seem to be some mechanism whereby the waves might be strongly beamed in the galactic plane. The sun is close to the galactic plane. With such beaming it could be that Weber's detectors are picking up a high proportion of the energy emitted at the galactic center. If that is the case, all the observational conflicts could be removed. Attempts have been made to explain such beaming on the basis of the existence of a large, strongly rotating black hole at the galactic center, but so far these attempts have not been very convincing.

It may be that no explanation based on black holes can be made to work. If Weber's observations remain, then rather than abandon the general theory of relativity (except in regions of large space-time curvature where we might expect the classical theory to break down anyway) we may prefer to seek an explanation in terms of naked singularities. It is worth pointing out that some of Kerr's solutions of Einstein's equations embody naked singularities and that Carter has calculated that any effect originating at a naked singularity can be observed in only one plane. If this phenomenon somehow applied to the center of our galaxy, we could imagine the singularity's being responsible both for the existence of the galactic plane and for Weber's gravitational waves. This would remove the observational discrepancies, although it would be a radical explanation of Weber's observations. In comparison black holes can now be regarded as being "conventional"; indeed, for this reason they must be preferred a priori in any attempt at explanation. Yet nature does not always prefer conventional explanations, least of all in astronomy.

52

TWISTOR THEORY: AN APPROACH TO THE QUANTISATION OF FIELDS AND SPACE-TIME

Phys. Repts. **6C** (1972), no. 4, 241–315. With M. A. H. MacCallum

Many significant twistor ideas are introduced in this article. The interpretation of a twistor in terms of momentum and angular momentum is made explicit, as are the quantum commutation rules (see Chapter 46). Asymptotic twistors are also introduced. Of considerable current interest are the twistor diagrams (the twistor analogues of the Feynman diagrams of quantum field theory) put forward here for the first time. The theory of twistor diagrams has been subsequently developed into a powerful calculus for particle scatterings by Andrew Hodges.

430 Twistor theory: An approach to the quantisation of fields and space-time. *Phys. Repts.* **6C** (1972), no. 4, 241–315

R. PENROSE

Birkbeck College, London, England

and

M.A.H. MacCALLUM

King's College, Cambridge, England

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM

Twistor theory: An approach to the quantisation of fields and space-time. *Phys. Repts.* **6C** (1972), no. 4, 241–315

TWISTOR THEORY: AN APPROACH TO THE QUANTISATION OF FIELDS AND SPACE-TIME

R. PENROSE
Birkbeck College, London, England

and

M.A.H. MacCALLUM
King's College, Cambridge, England

Received 9 June 1972

Contents:

Preface 243
0. Introduction 243
1. Preliminaries 245
  1.1. Conformal transformations 245
  1.2. Spinors 248
  1.3. Momentum and angular momentum 253
2. Twistors in flat space 256
  2.1. Basic ideas; conformal invariance 256
  2.2. Twistor space and Minkowski geometry 258
  2.3. Solutions of the zero-rest-mass field equations 261
  2.4. Quantisation 264
3. Twistors in curved space 266
  3.1. Local twistors 267
  3.2. Global twistors 269
  3.3. Quantisation 274
4. Evaluation of scattering amplitudes 276
  4.1. Introduction 276
  4.2. Rules for diagrams 282
  4.3. Evaluation of diagrams using elementary states 286
  4.4. Momentum state diagrams 292
5. Further developments 298
  5.1. Asymptotic twistors 298
  5.2. Scattering off a fixed source 307
  5.3. Potentials and massive fields 312
References 314

Abstract:

Twistor theory offers a new approach, starting with conformally-invariant concepts, to the synthesis of quantum theory and relativity. Twistors for flat space-time are the SU(2,2) spinors of the twofold covering group O(2,4) of the conformal group. They describe the momentum and angular momentum structure of zero-rest-mass particles. Space-time points arise as secondary concepts corresponding to linear sets in twistor space. They, rather than the null cones, should become "smeared out" on passage to a quantised gravitational theory. Twistors are represented here in two-component spinor terms. Zero-rest-mass fields are described by holomorphic functions on twistor space, on which there is a natural canonical structure leading to a natural choice of canonical quantum operators. The generalisation to curved space can be accomplished in three ways; i) local twistors, a conformally invariant calculus, ii) global twistors, and iii) asymptotic twistors which provide the framework for an S-matrix approach in asymptotically flat space-times. A Hamiltonian scattering theory of global twistors is used to calculate scattering cross-sections. This leads to twistor analogues of Feynman graphs for the treatment of massless quantum electrodynamics. The recent development of methods for dealing with massive (conformal symmetry breaking) sources and fields is briefly reviewed.

Single orders for this issue

PHYSICS REPORTS (Section C of PHYSICS LETTERS) 6, no. 4 (1972) 241–316.

Copies of this issue may be obtained at the price given below. All orders should be sent directly to the Publisher. Orders must be accompanied by check.

Single issue price Hfl. 25.– ($ 8.00), postage included.

432 Twistor theory: An approach to the quantisation of fields and space-time. *Phys. Repts.* **6C** (1972), no. 4, 241–315

Preface

These notes are an extended, revised and edited version of M. MacCallum's notes of a course given by R. Penrose at the Institute of Theoretical Astronomy, Cambridge, during the spring and summer of 1970. Additional material by R. Penrose on subsequent developments appears as a concluding section. Financial support for the lectures was provided by the Institute of Theoretical Astronomy.

0. Introduction

It is well known that there are a number of unsatisfactory features of our present ideas about physics. Among these are the infinite divergences of quantum field theory, the lack of a really convincing synthesis of quantum theory and general relativity, and perhaps also our dependence upon the notion of a continuum without any real physical evidence. Twistor theory is an attempt at a new formalism for the description of basic physical processes which has relevance to these problems and it is hoped that when the theory becomes more complete a new outlook on them will be provided. If the attempt is successful, it would of course have very wide implications for all of physics. For everyday purposes our present theories would naturally suffice but our viewpoint would be changed just as the development of relativity modified our view of Newtonian mechanics. Although no final assessment of twistor theory's success can yet be made the results have been sufficiently encouraging for us to feel it worthwhile preparing a reasonably up to date and unified account for the use of colleagues in different branches of physics.

The last two of the difficulties mentioned above are clearly related. If space-time is no longer regarded as a continuum, it will no longer be valid to think of either the quantum fields or the gravitational fields in the usual way. One can in fact argue [38] that to accept that there are as many points in 10^{-13} cm or even 10^{-1000} cm as there are in the entire universe is physically unrealistic and that our use of the continuum arises solely from its mathematical utility. We take the view that to encompass quantum theory and general relativity satisfactorily one needs to do more than simply apply some suitable quantisation technique to solutions of Einstein's equations. One should rather be thinking of quantising space-time itself. This should not be conceived as simply replacing the continuum by a discrete set of points (though this has been attempted) but rather as seeking a way of treating points as "smeared out" just as quantum theory smears out particles.

In earlier work [33, 37] it was shown that one could build up the notion of the Euclidean space from the limit of the interaction probabilities of a large network of particles quasi-statically exchanging spin. The Euclidean structure in this development arises from the combinatorial rules satisfied by total angular momentum in non-relativistic quantum mechanics. In the same way that SU(2) spinors provide a basis for the description of non-relativistic angular momentum, twistor theory can be used to describe relativistic angular momentum in a unified way, the concepts of spin and orbital angular momentum uniting together appropriately. The hope is that developments of the twistor picture will eventually enable us to construct Lorentzian manifolds to serve as models of space-time. Certainly points of space-time are dependent quantities in the twistor formalism, the twistors themselves playing the basic role. However the complex continuum still plays a large part. Indeed the complex numbers and holomorphic functions which are already basic to modern particle physics now appear mixed up with the structure of space-time itself. Nevertheless, holomorphic functions have a certain "rigidity" suggestive of a possible underlying combinatorial structure.

The twistor theory is in fact largely based on ideas of conformal invariance, zero rest mass particles and conformally invariant fields being taken as a fundamental aspect of important parts of physics. In this respect twistor theory has a connection with current work by particle physicists, who have been exploring the implications of conformal invariance with considerable vigour [34, 39].

A twistor (of the simplest type) can be pictured "classically" as effectively a zero rest-mass particle in free motion, where the particle may possess an intrinsic spin, and also a "phase" which can be realized as a kind of polarization plane. Such twistors form an eight-real-dimensional manifold, which can be described in a natural way as a vector space of four complex dimensions. This vector space (twistor space) in effect replaces the space-time as the background in terms of which physical phenomena are to be described. Space-time points can then be reconstructed from the twistor space (being represented as certain linear subspaces), but they become secondary to the twistors themselves. Furthermore, when general relativity and quantum theory both become involved, it is to be expected that the concept of a space-time point should cease to have precise meaning within the theory. In effect, the space-time points become "smeared" by the uncertainty principle (rather than the light cones becoming "smeared" and the points not – which has been a more usual viewpoint).

As the theory stands it does not provide a formulation of a quantised general relativity; quantized gravitational interactions have not yet been successfully incorporated into the theory. Nor has a full treatment of particles with non-vanishing rest-mass emerged. On the other hand, the theory appears to give correct answers for scattering processes involving massless charged particles and photons (i.e. high energy limit of quantum electrodynamics) and it may even yield some new insights into the nature of the electromagnetic interaction. The theory so far appears to be successful in avoiding divergences, i.e. the calculations that have been carried out do not lead to infinities in the same way as does the conventional formalism and it seems that such infinities should be absent altogether. It is hoped that when the theory becomes more complete, this feature will be retained.

The difficulties confronting the theory in respect of gravitational interactions and rest-mass appear to be related to the fact that these are things which *break conformal invariance*. Massless particles and electromagnetic interactions, on the other hand, *are* conformally invariant concepts. The basic formalism exhibits manifest conformal invariance, so if conformal invariance is to be broken, this must apparently be done explicitly, with the aid of auxiliary elements which do not share in the invariance. One possible method of incorporating such elements is suggested by the work described in section 5.3. The essential act of faith on which the utility of the twistor formalism depends, is that it should be useful to isolate the conformally non-invariant aspects of physics from the conformally invariant ones, and that having done this, a large and important body of physical processes will be seen as possessing conformal invariance.

Twistors (that is to say, the original flat-space twistors about which these notes are mainly concerned) are actually the reduced spinors for the proper pseudo-orthogonal group SO(2,4) which is locally isomorphic with, and 2-1 homomorphic with, the restricted conformal group of flat space-time. They form a representation space for the pseudo-unitary group SU(2,2), this in turn being locally isomorphic and 2-1 homomorphic with SO(2,4). Thus, the simplest twistors are four-valued objects with four complex components, which are acted upon by the 15 parameter conformal group of flat space-time. The four-valuedness of twistors has not yet played any very important role in this theory, however.

One of the most striking features of twistor theory is the way in which complex numbers and holomorphic (i.e. complex analytic) structure emerge as concepts intimately involved in the geometry of space-time. We have become accustomed to the very basic role which complex numbers and holomorphic functions play in quantum theory, particularly that of elementary particles. It seems, therefore, that complex numbers are (at least at our present level of understanding) a very important constituent of the structure of physical laws. The twistor theory carries this further in suggesting that complex numbers may also be very basically involved in defining the nature of space-time itself. In addition, we shall see that the zero rest-mass field equations for each spin (wave equation, Weyl neutrino equation, Maxwell free-field equations, linearized Einstein equations) all emerge in a very simple way from the complex structure of twistor space, being obtained as contour integrals of holomorphic functions of twistors. The twistor picture "geometrises" the usual splitting of field amplitudes into positive and negative frequency parts by describing this in terms of the position of singularities of holomorphic functions. Thus the twistor formalism has the effect of uniting various aspects of the role, both quantum and classical, that complex numbers seem to play in physics.

The present notes should be regarded as to some extent provisional. Many problems remain to be solved in the theory. Even the difficulties involved in merely translating between the twistor formalism and the conventional formalism constitute a serious stumbling block. The twistor scattering diagrams described in section 4 do not always appear to be directly translatable into conventional (Feynman) terms and this leads to difficulties in interpretation. One must proceed to some extent by guesswork, but here severe problems of actually computing the twistor diagrams constitute another stumbling block. Nevertheless, despite these difficulties, we feel that some new insights into the nature of physical processes may possibly be discernable even in the theory as it stands. For example, if the twistor diagrams of section 4 are to be taken seriously from the physical point of view – and it is tempting to think that they can be – then there may be some significance in the twistor lines representing a kind of "half particle" which can be exchanged in virtual processes. This would appear to be related to the fact that a twistor is really a kind of "square root" of the structure of a zero rest-mass particle.

1. Preliminaries

1.1. *Conformal transformations*

There is a certain confusion in the literature owing to the fact that two quite distinct concepts are both given the name "conformal transformation".

The first of these, which we shall refer to (cf. [34]) as a *conformal rescaling*, consists solely of a replacement[1]

$$g_{ab} \to \hat{g}_{ab} = \Omega^2 g_{ab} \tag{1.1}$$

---

[1] See footnote on the next page.

of the space-time metric g_{ab} by a conformally-related one $\hat{g}_{ab}$, Ω being a smooth positive scalar field on the underlying manifold. Thus the interval ds is transformed to $d\hat{s} = \Omega ds$. If g_{ab} is a flat metric, then g_{ab} will in general not be flat, though it will of course be conformally flat. The conformal rescalings of a given space-time form an infinite-parameter Abelian group. The points of the space-time are unaffected by a conformal rescaling. The null cones, and, therefore, the causal structure of the space-time, are also unchanged.

The other type of conformal transformation is what we shall call a *conformal mapping*. This is a smooth mapping μ which carries each point of a space-time N to a point of some space-time $\tilde{\mathrm{N}}$ in such a way that the metric on $\tilde{\mathrm{N}}$ induced by μ from that of N is a conformal rescaling of the original metric on $\tilde{\mathrm{N}}$. In other words the map μ preserves null cone structure.

Conformal mappings of Minkowski space[2] M′ to itself have a particular interest. These include the Poincaré transformations, which are metric-preserving, and the simple overall dilations, whose corresponding rescaling multiplies the metric at each point by a constant factor. The remainder are generated by the (involutary) inversions

$$\hat{x}^a = -x^a(x^b x_b)^{-1}, \qquad x^a = -\hat{x}^a(\hat{x}^b \hat{x}_b)^{-1} \tag{1.2}$$

which are a 4-parameter set since the choice of origin is arbitrary. These transformations preserve the time sense but involve spatial reflection. They are conformal mappings since the induced and original metrics are related by

$$d\hat{s}^2 = d\hat{x}^a d\hat{x}_a = dx^a dx_a/(x^b x_b)^2 = \Omega^2 ds^2. \tag{1.3}$$

However, these transformations do not involve only the points of M′, because the null cone of the origin is sent to infinity. We therefore introduce compactified Minkowski space M, which consists of M′ together with a closed null cone at infinity. We may picture the structure in terms of two cones joined base to base, the interior being M′ and the two bounding cones being identified along opposite generators with future sense preserved (see fig. 1). (Thus the "equator" I_0 must be considered as a single point.) For fuller discussion of the structure of M, see [2–4]. Note that one can consider the equations (1.2) as expressing a coordinate change, rather than a point transformation, on M; and that the null cone at infinity is on the same footing as any other null cone in M as far as conformal mapping symmetry is concerned (see [3]). In consequence of the latter, the concept of radiation is not conformally invariant, since it depends on knowing where infinity is.

The conformal mapping group of M is of 15 parameters (and non-abelian). We shall here concern ourselves with the restricted conformal group, i.e. the subgroup of mappings connected with the

---

[1] Footnote from preceding page. We adopt throughout these notes the "abstract index" convention [1, 2]. Abstract indices are non-numerical and serve only as organisational markers enabling the basically coordinate-free operations of contraction, index-permutation, outer multiplication, addition and complex conjugation to be expressed in a transparent, yet frame-independent, way. Latin and Greek indices will be used in this abstract way, while Gothic and Hebrew indices will be used for the corresponding "normal" indices (i.e. to represent the components in some particular frame). Latin lower case letters will be abstract tensor indices, upper case Latin letters abstract spinor indices and Greek letters abstract twistor indices. Thus Gothic lower case (tensor) and Hebrew (twistor) letters will range from 0 to 3, while Gothic upper case (spinor) letters range from 0 to 1. In all cases round brackets surrounding indices denote symmetrisation, square brackets skew-symmetrisation.

[2] Minkowski space is simply flat space-time ("space-time" meaning a pseudo-Riemannian Hausdorff manifold of signature −2 and dimension 4). Coordinates x^a used without definition are the natural pseudo-Cartesian ("flat") ones. We can also use x^a with an abstract index to denote the position vector with respect to some origin. Also ∇_a denotes the covariant derivative (or gradient) so $\nabla_a x^b = \delta_a{}^b$. Thus in flat coordinates $\nabla_a = \partial/\partial x^a$.

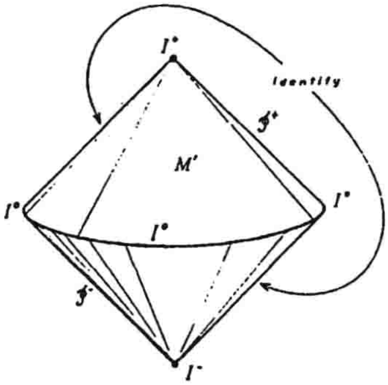


Fig. 1. Compactified Minkowski space M. I^0, I^+, I^- are points at spatial infinity, future time infinity and past time infinity respectively, while I^+, I^- are future and past null infinity cones (cf. [3]). The compactified space has I^0, I^+, I^- identified and I^-, I^+ identified along opposite generators. For typographical reasons, "I" replaces the more usual script I depicted in figs. 1 and 40.

identity map[3]. This does not include the actual mappings (1.2) but does include their products with space reflections. It is 2-1 covered by (and so locally isomorphic with) the six-dimensional pseudo-orthogonal group SO(2,4) which in turn is 2-1 covered by SU(2,2), a group of unimodular pseudo-unitary matrices [6].

The infinitesimal conformal motions[4] are described by the conformal Killing vectors ξ^a and are given, for infinitesimal ϵ, by

$$x^a \to x^a + \epsilon\xi^a.$$

The vectors ξ_a must satisfy

$$\nabla_{(a}\xi_{b)} = \tfrac{1}{4}g_{ab}(\nabla_c\xi^c). \tag{1.4}$$

The general solution of this is

$$\xi_a = S_{ab}x^b + T_a + Q_a(x^c x_c) - 2x_a(x^c Q_c) + Rx_a \tag{1.5}$$

where $S_{ab} = S_{[ab]}$ generate the Lorentz rotations (6 parameters), T_a the translations (4 parameters), R the dilations (1 parameter) and Q_a the so-called "uniform acceleration" transformations [8] (4 parameters). (This terminology is rather misleading, however, and will be avoided here. A more correct use of the terminology "uniform acceleration" is for a coordinate transformation which makes the Minkowski metric take the form

$$\mathrm{d}s^2 = z^2\mathrm{d}t^2 - \mathrm{d}x^2 - \mathrm{d}y^2 - \mathrm{d}z^2 \qquad \text{cf. [2].)}$$

Questions of conformal invariance are handled most easily within the framework of conformal rescalings rather than conformal mappings. A physical theory will be said to be *conformally invariant* if it is possible to attach conformal weights to all the quantities appearing in the theory in such a way that all field equations are preserved under conformal rescalings. (A tensor or spinor $A^{\cdots}{}_{\cdots}$ is said to have conformal weight r if we are to make the replacement $A^{\cdots}{}_{\cdots} \rightarrow \hat{A}^{\cdots}{}_{\cdots} = \Omega^r A^{\cdots}{}_{\cdots}$ under the conformal rescaling $g_{ab} \rightarrow \hat{g}_{ab} = \Omega^2 g_{ab}$.) A flat space theory which is Poincaré invariant and also conformally invariant in this sense, will then be invariant under the 15-parameter conformal group. This is because the Poincaré motions of Minkowski space become conformal motions according to any other conformally rescaled flat metric. Conformal motions obtainable in this way are sufficient to generate the full conformal group. But the type of conformal invariance described above is really more general than this since the conformal rescalings need not apply to flat space-time at all or even to conformally flat space-times.

---

[3] Thus we will not be concerned with reflections. For some discussion of reflections in twistor theory see [5].

[4] For an introduction to the description of Lie groups of transformations of a manifold in terms of the generating vectors (with which physicists accustomed to considering Lie groups only in terms of their representations may be unfamiliar) see e.g. [7].

In order to establish conformal invariance of a theory, one needs to know how to transform the (covariant) derivative operator under conformal rescaling. Remarkably enough, this is rather simpler within the two-component spinor formalism than within the tensor formalism. Since two-component spinors will also play an essential role in other aspects of twistor theory, we will next briefly summarise the relevant notation and methods.

1.2. *Spinors*

The essential fact on which the 2-component spinor calculus is based is the local isomorphism between the Lorentz group and the group SL(2, ℂ) of complex unimodular 2 × 2 matrices (which is the covering group of the identity-connected component of the Lorentz group). It should be noted that this does not mean that spinors can only be used in flat space, since it is possible to use this isomorphism locally is curved space-time [9, 40].

Representing the Minkowksi components $u^{\mathfrak{a}}$ of a world vector u^a according to the matrix scheme

$$u^{\mathfrak{a}} = (u^0, u^1, u^2, u^3) \leftrightarrow u^{\mathfrak{A}\mathfrak{A}'} = \begin{pmatrix} u^{00'} & u^{01'} \\ u^{10'} & u^{11'} \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} u^0 + u^1 & u^2 + iu^3 \\ u^2 - iu^3 & u^0 - u^1 \end{pmatrix} \tag{1.6}$$

we find that when the components $u^{\mathfrak{a}}$ undergo a restricted Lorentz transformation L the $u^{\mathfrak{A}\mathfrak{A}'}$ undergo

$$u^{\mathfrak{A}\mathfrak{A}'} \rightarrow \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \begin{pmatrix} u^{00'} & u^{01'} \\ u^{10'} & u^{11'} \end{pmatrix} \begin{pmatrix} \bar{\alpha} & \bar{\gamma} \\ \bar{\beta} & \bar{\delta} \end{pmatrix} \tag{1.7}$$

where $\alpha, \beta, \gamma, \delta$ are complex and their matrix has unit determinant, i.e. $\alpha\delta - \beta\gamma = 1$. The hermiticity of $u^{\mathfrak{A}\mathfrak{A}'}$, i.e. reality of $u^{\mathfrak{a}}$, is preserved and so is

$$\{(u^0)^2 - (u^1)^2 - (u^2)^2 - (u^3)^2\} = 2 \det (u^{\mathfrak{A}\mathfrak{A}'}). \tag{1.8}$$

We can express (1.7) as

$$u^{\mathfrak{A}\mathfrak{A}'} \rightarrow S(\mathrm{L})^{\mathfrak{A}}{}_{\mathfrak{B}} u^{\mathfrak{B}\mathfrak{B}'} \overline{S(\mathrm{L})}^{\mathfrak{A}'}{}_{\mathfrak{B}'} \tag{1.9}$$

where $S(\mathrm{L}) \in \mathrm{SL}(2, \mathbb{C})$. Primed and unprimed indices must be treated as essentially different as regards contractions and permutations, but they are related to each other by complex conjugation[5], which converts a primed index into an unprimed one and vice versa.

---

[5] I.e. unprimed capital Gothic superfixes represent components transforming under the representation $S(\mathrm{L})$; primed superfixes under the complex conjugate representation $\overline{S(\mathrm{L})}$; and primed and unprimed suffixes under the transposed inverses of $S(\mathrm{L}), \overline{S(\mathrm{L})}$ respectively.

CHAPTER 52

The correspondence (1.6) shows how one may relate tensor and spinor components according to a standard scheme, but there is nothing special about this particular correspondence. From the point of view of the abstract index notation[1], the essential feature is that each abstract tensor index (four-dimensional) is to be equated with a pair of (two-dimensional) spinor indices, one primed and one unprimed. Thus the abstract tensor indices $a, b, c\ldots, a_0, b_0\ldots, a_1\ldots,$ may be expressed as $a = AA', b = BB', \ldots a_0 = A_0A_0'\ldots$ and we can write

$$u^a = u^{AA'}.$$

(The reader who prefers to retain a component description such as that of (1.6) can re-express our equations in these terms by use of the Infeld-van der Waerden symbols $\sigma^a{}_{\mathfrak{A}\mathfrak{A}'}$, $\sigma_a{}^{\mathfrak{A}\mathfrak{A}'}$ [40].) Also (cf. 1.8))

$$g_{ab} = \epsilon_{AB}\epsilon_{A'B'}, \qquad g^{ab} = \epsilon^{AB}\epsilon^{A'B'} \tag{1.10}$$

where the ϵ's are skew-symmetrical with $\epsilon_{A'B'} = \overline{\epsilon_{AB}}$, $\epsilon^{A'B'} = \overline{\epsilon^{AB}}$ (i.e. their coordinate representations under (1.6) are $\left(\begin{smallmatrix}0 & 1\\ -1 & 0\end{smallmatrix}\right)$). We use ϵ's to raise and lower indices, thus

$$\xi_B = \xi^A\epsilon_{AB}; \qquad \xi^A = \epsilon^{AB}\xi_B; \qquad \eta_{B'} = \eta^{A'}\epsilon_{A'B'}; \qquad \eta^{A'} = \epsilon^{A'B'}\eta_{B'}. \tag{1.11}$$

The tensor and spinor "Kronecker deltas" will be written $g_b{}^a$ and $\epsilon_A{}^B$, $\epsilon_{A'}{}^{B'}$, respectively. Thus

$$g_a{}^b = \epsilon_A{}^B\epsilon_{A'}{}^{B'}$$

and

$$\chi^{\cdots a}{}_{\cdot\cdot}{}^{\cdot}\, g_a{}^b = \chi^{\cdot\cdot b}{}_{\cdot\cdot}{}^{\cdot}, \qquad \psi_{\ldots A\ldots}\epsilon_B{}^A = \psi_{\ldots B\ldots}, \qquad \theta_{\cdot\cdot}{}^{\cdot A'}{}_{\cdot\cdot}{}^{\cdot}\,\epsilon_{A'}{}^{B'} = \theta_{\cdot\cdot}{}^{\cdot B'}{}_{\cdot\cdot}{}^{\cdot},$$

etc.. A complex null vector $u^a(u_bu^b = 0)$ has a spinor form

$$u^{AA'} = \xi^A\eta^{A'}$$

(since the matrix of components $u^{\mathfrak{A}\mathfrak{A}'}$ has rank $\leqslant 1$, cf. (1.8)[6]). If u^a is real, then[7]

$$u^{AA'} = \pm\, \xi^A\bar{\xi}^{A'}. \tag{1.12}$$

The plus sign occurs if u^a is future-pointing and the minus if u^a is past-pointing. Note that $\xi^A\xi_A = 0$ (as ϵ_{AB} is skew) so that $u^au_a = 0$ follows directly from (1.12). Conversely if $\xi^A\zeta_A = 0$ then ξ^A is a scalar multiple of ζ^A (or $\zeta_A = 0$).

A spinor ξ_A contains more information than the corresponding null vector given by (1.12). A non-zero spinor has a geometrical interpretation, up to an essential sign ambiguity, as a *null flag* [2, 10]. This consists of the corresponding null vector u^a ("flagpole") and a null 2-plane ("flag

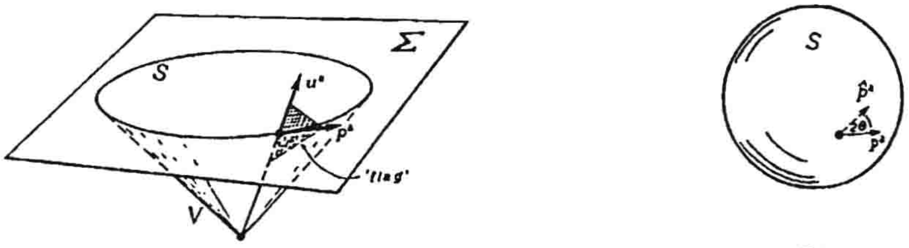
---

[6] This, and other simplifying features, arise from the two-dimensionality of the vector space of spinors. As another example, any two-spinor α_{AB} can be expressed as a sum of a symmetrised outer-product of one-index spinors and a skewed outer-product.

[7] Here we will write conjugate of a quantity as

$$\overline{\chi^{A'}{}_{BC'}{}^{D'}} = \bar{\chi}^{A}{}_{B'C}{}^{D'}$$

(except in the case of $\epsilon_{A'B'}$ etc. and basis spinors o^A, ι^B where the bar will be omitted). Many authors omit the bar on the right-hand symbol but we retain it for the sake of clarity and notational consistency and in addition because of the flexibility it allows in the placing of indices, i.e. we can write $\phi_{AA'BB'} = \phi_{ABA'B'} = \phi_{A'B'AB}$ without confusion while $\bar{\phi}_{AA'BB'} = \bar{\phi}_{A'B'AB} = \bar{\phi}_{ABA'B'}$.



(a) One spatial dimension suppressed.

(b) Time dimension suppressed.

Fig. 2. (a) The spinor ξ_A defines a "null flag". This may be pictured as a polarisation vector tangent to the "celestial sphere" S. (b) shows how p^b is rotated when the phase of ξ_A is altered.

plane") element which contains and is orthogonal to the flagpole. This latter is defined by the bivector

$$F_{ab} = u_{[a}p_{b]} = \epsilon_{AB}\xi_{A'}\xi_{B'} + \epsilon_{A'B'}\xi_A\xi_B, \tag{1.13}$$

where $p_a = 2(\xi_A\bar{\lambda}_{A'} + \lambda_A\xi_{A'})$ for some λ_A with $\xi_A\lambda^A = 1$. When the phase of ξ_A is altered (i.e. $\xi_A \to e^{i\theta}\xi_A$) the vector u_a is unaltered, while p_b turns through an angle 2θ. We may consider a spacelike hyper-plane Σ intersecting the null cone V, of a point 0, in S (a 2-sphere). ξ_A then describes a point on S, and a vector tangent to S which defines a polarisation direction. As θ varies this polarisation vector sweeps out the 2-plane tangent to S, and performs one revolution through 2π as θ changes by π (see fig. 2).

Our (covariant) derivative operator $\nabla_a = \nabla_{AA'}$ satisfies

$$\nabla_a\epsilon_{BC} = 0, \qquad \nabla_a\epsilon_{B'C'} = 0$$

(whence $\nabla_a g_{bc} = 0$), and

$$\nabla_a\nabla_b\phi = \nabla_b\nabla_a\phi.$$

In curved space-time, we have the relation

$$(\nabla_{AA'}\nabla_{BB'} - \nabla_{BB'}\nabla_{AA'})\xi_C = \Psi_{ABCD}\xi^D\epsilon_{A'B'} - 2\Lambda\xi_{(A}\epsilon_{B)C}\epsilon_{A'B'} + \Phi_{CDA'B'}\xi^D\epsilon_{AB}$$

where the curvature spinors Ψ_{ABCD}, $\Phi_{ABC'D'}$, Λ have the symmetries

$$\Psi_{ABCD} = \Psi_{(ABCD)}, \qquad \Phi_{ABC'D'} = \bar{\Phi}_{ABC'D'} = \Phi_{(AB)(C'D')}, \qquad \Lambda = \bar{\Lambda} \tag{1.15}$$

and are related to the curvature tensor R_{abcd} (with sign convention $(\nabla_a\nabla_b - \nabla_b\nabla_a)V_d = R_{abcd}V^c$) by

$$R_{abcd} = \Psi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\bar{\Psi}_{A'B'C'D'} + 2\Lambda(\epsilon_{AC}\epsilon_{BD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\epsilon_{A'D'}\epsilon_{B'C'}) + \epsilon_{A'B'}\Phi_{ABC'D'}\epsilon_{CD} + \\ + \epsilon_{AB}\Phi_{CDA'B'}\epsilon_{C'D'}. \tag{1.16}$$

We thus have

$$\Phi_{ABA'B'} = -\tfrac{1}{2}R_{ab} + \tfrac{1}{8}Rg_{ab}, \qquad \Lambda = R/24$$

(where $R_{ab} = R^{c}{}_{acb}$, $R = R^{a}{}_{a}$) and

$$\Psi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\overline{\Psi}_{A'B'C'D'} = C_{abcd}$$

where C_{abcd} is Weyl's conformal curvature tensor, whose vanishing is a necessary and sufficient condition for the space-time to be conformally flat.

We also introduce, for future use, the completely skew tensor

$$\eta^{abcd} = \eta^{[abcd]}$$

defined by $\eta^{0123} = (\sqrt{-g})^{-1}$, so that $\eta_{0123} = -\sqrt{-g}$, in a right-handed coordinate system. Its spinor equivalent is given by

$$\eta_{abcd} = \mathrm{i}\,\epsilon_{AD}\epsilon_{BC}\epsilon_{A'C'}\epsilon_{B'D'} - \mathrm{i}\,\epsilon_{AC}\epsilon_{BD}\epsilon_{A'D'}\epsilon_{B'C'}\,. \tag{1.17}$$

Now we can discuss how conformal rescalings affect spinors. Under the rescaling $g_{ab} \to \hat{g}_{ab} = \Omega^2 g_{ab}$ we take

$$\hat{\epsilon}_{AB} = \Omega\epsilon_{AB}, \qquad \hat{\epsilon}_{A'B'} = \Omega\epsilon_{A'B'}\ , \qquad \hat{\epsilon}^{AB} = \Omega^{-1}\epsilon^{AB}, \qquad \hat{\epsilon}^{A'B'} = \Omega^{-1}\epsilon^{A'B'}. \tag{1.18}$$

We have $\hat{\nabla}_a\phi = \nabla_a\phi$ when ∇_a acts on a scalar. When ∇_a acts on spinors we have

$$\begin{aligned} &\hat{\nabla}_{AA'}\xi_B = \nabla_{AA'}\xi_B - \Upsilon_{BA'}\xi_A, && \hat{\nabla}_{AA'}\eta_{B'} = \nabla_{AA'}\eta_{B'} - \Upsilon_{AB'}\eta_{A'} \\ &\hat{\nabla}_{AA'}\xi^B = \nabla_{AA'}\xi^B + \epsilon_A{}^B\Upsilon_{CA'}\xi^C, && \hat{\nabla}_{AA'}\eta^{B'} = \nabla_{AA'}\eta^{B'} + \epsilon_{A'}{}^{B'}\Upsilon_{AC'}\eta^{C'} \end{aligned} \tag{1.19a}$$

where

$$\Upsilon_a = \Omega^{-1}\nabla_a\Omega. \tag{1.19b}$$

When ∇_a acts on spinors of higher valence we simply treat each index in turn according to the above scheme (so there is one term involving Υ_a for each index). For example,

$$\begin{aligned} \hat{\nabla}_{CC'}\hat{\epsilon}_{AB} &= \hat{\nabla}_{CC'}(\Omega\epsilon_{AB}) \\ &= (\nabla_{CC'}\Omega)\epsilon_{AB} + \Omega(\nabla_{CC'}\epsilon_{AB}) + (\nabla_{AC'}\Omega)\epsilon_{CB} + (\nabla_{BC'}\Omega)\epsilon_{AC} \\ &= 3\,\epsilon_{[AB}\nabla_{C]C'}\Omega = 0 \end{aligned}$$

(by the 2-dimensionality of spinor space). The covariant derivative of a vector transforms as

$$\begin{aligned} \hat{\nabla}_a V_b = \hat{\nabla}_{AA'}V_{BB'} &= \nabla_{AA'}V_{BB'} - \Upsilon_{BA'}V_{AB'} - \Upsilon_{AB'}V_{BA'} \\ &= \nabla_a V_b - \Upsilon_a V_b - \Upsilon_b V_a + (\Upsilon_{AA'}V_{BB'} + \Upsilon_{BB'}V_{AA'} - \Upsilon_{BA'}V_{AB'} - \Upsilon_{AB'}V_{BA'}) \\ &= \nabla_a V_b - \Upsilon_a V_b - \Upsilon_b V_a + \epsilon_{AB}\epsilon_{A'B'}\Upsilon_{CC'}V^{CC'} \\ &= \nabla_a V_b - \Upsilon_a V_b - \Upsilon_b V_a + g_{ab}(\Upsilon_c V^c) \end{aligned} \tag{1.20}$$

(using $\chi_{\ldots AB\ldots} - \chi_{\ldots BA\ldots} = \epsilon_{AB}\epsilon^{CD}\chi_{\ldots CD\ldots}$). Note that this generates the transform under conformal transformations of ∇_a applied to tensor indices, as (1.19) does for spinors.

Using this information we find the following transformation laws for the curvature

$$\hat{\Psi}_{ABCD} = \Psi_{ABCD}$$

and

$$\hat{P}_{ABC'D'} = P_{ABC'D'} - \nabla_{AC'}\Upsilon_{BD'} + \Upsilon_{AD'}\Upsilon_{BC'} \tag{1.21}$$

where

$$P_{ABA'B'} = \Phi_{ABA'B'} - \Lambda\epsilon_{AB}\epsilon_{A'B'} = \tfrac{1}{12}Rg_{ab} - \tfrac{1}{2}R_{ab}.$$

The Bianchi identities $\nabla_{[a}R_{bc]de} = 0$, which are equivalent to

$$\nabla^a C_{abcd} = 2\,\nabla_{[c}P_{d]b},$$

become

$$\nabla^D_{A'}\Psi_{ABCD} = -\nabla^{B'}_{(C}P_{A)BA'B'} \tag{1.22}$$

which in empty space-time ($P_{ab} = R_{ab} = 0$) simplifies to

$$\nabla^{AD'}\Psi_{ABCD} = 0. \tag{1.23}$$

Finally let us consider some conformally-invariant theories. For example, Maxwell's equations

$$\nabla_a F^{ba} = 4\pi\, J^b \text{ and } \nabla_{[a}F_{bc]} = 0$$

are conformally invariant if we set $\hat{F}_{ab} = F_{ab}$ and $\hat{J}_a = \Omega^{-2}J_a$ (so $\hat{F}^{ab} = \Omega^{-4}F^{ab}; \hat{J}^a = \Omega^{-4}J^a$). That is to say we get

$$\hat{\nabla}_a\hat{F}^{ba} = 4\pi\,\hat{J}^b \text{ and } \hat{\nabla}_{[a}\hat{F}_{bc]} = 0.$$

This may be verified in various ways, e.g. by using the tensor formula for $\hat{\nabla}_a$ above, or by using the spinor formulae applied to the spinor version

$$\nabla^A_{B'}\phi_{AB} = -2\pi J_{BB'} \tag{1.24}$$

of Maxwell's equations, where

$$F_{ab} = \phi_{AB}\epsilon_{A'B'} + \epsilon_{AB}\bar{\phi}_{A'B'}$$

with $\phi_{AB} = \phi_{(AB)}$, $\hat{\phi}_{AB} = \Omega^{-1}\phi_{AB}$. When $J_a = 0$, (1.24) becomes a particular case of the zero-rest-mass free-field equations for spin $\frac{1}{2}n$

$$\nabla^{AP'}\phi_{AB\ldots L} = 0 \tag{1.25}$$

where $\phi_{AB\ldots L}$ is symmetric with n indices (see [12]). If $n = 0$ we adopt the second-order equation

$$(\nabla_a\nabla^a + \tfrac{1}{6}R)\phi = 0. \tag{1.26}$$

For each n, these equations are conformally invariant if $\hat{\phi}_{AB\ldots L} = \Omega^{-1}\phi_{AB\ldots L}$ as is readily verified using (1.19) etc. [13]. The case $n = 4$ has particular interest since the tensor

$$K_{abcd} = \phi_{ABCD}\epsilon_{A'B'}\epsilon_{C'D'} + \epsilon_{AB}\epsilon_{CD}\bar{\phi}_{A'B'C'D'}$$

defined from a solution of (1.25) in Minkowski space, satisfies

$$\begin{aligned} &K_{abcd} = K_{[ab][cd]} = K_{cdab}, && K_{a[bcd]} = 0, \\ &K^{a}{}_{bcd} = 0, \quad \nabla^a K_{abcd} = 0 && (\text{whence } \nabla_{[a}K_{bc]de} = 0) \end{aligned} \tag{1.27}$$

and represents a source-free gravitational field in the linearised theory. In this it is assumed that

$$g_{ab} = \eta_{ab} + \epsilon h_{ab}$$

where η_{ab} is the Minkowski metric, ϵ is infinitesimal and h_{ab} is some symmetric tensor. K_{abcd} is locally[*] the Weyl (or in empty space, Riemann) tensor for some metric of this form, provided it satisfies (1.27). It should be noted that conformal rescaling of the metric gives $\hat{C}_{abcd} = \Omega^{-2} C_{abcd}$ while $\hat{K}_{abcd} = \Omega^{-1} K_{abcd}$. For further details of spinor calculus see [2, 9, 15, Pirani in 35].

Finally we note that to describe states in quantum mechanical systems, complex vectors and tensors are used. If the operator $i\hbar\, \partial/\partial t$ has positive eigenvalue, the quantum state has positive energy. We will describe these and the corresponding classical states as having positive frequency. It turns out the solutions of (1.25) with positive energy represent negative helicity particles and the positive energy solutions of the conjugate equation

$$\nabla^{AP'}\theta_{A'B'\ldots L'} = 0 \tag{1.28}$$

have the other helicity. This essential difference reappears in section 2.3. Raising and lowering of indices only alters the conformal weight, but complex conjugation of the spinor reverses the helicity.

For example a free photon wave function is described by a complex F_{ab}. When translated into spinor form this gives rise to independent spinors ϕ_{AB}, $\theta_{A'B'}$, by

$$F_{ab} = \epsilon_{A'B'}\phi_{AB} + \epsilon_{AB}\theta_{A'B'}$$

and ϕ_{AB}, $\theta_{A'B'}$ satisfy respectively (1.25) and (1.28). Considering a plane wave, we find

$$\phi_{AB} = \alpha_A\alpha_B \exp\{-\mathrm{i}(\alpha_P\bar{\alpha}_{P'}x^{PP'})\}$$

$$\theta_{A'B'} = \bar{\alpha}_{A'}\bar{\alpha}_{B'} \exp\{-\mathrm{i}(\alpha_P\bar{\alpha}_{P'}x^{PP'})\}\,.$$

The ϕ_{AB} thus derived corresponds to the ϕ_{AB} for a real circularly-polarised wave, which is in fact left-handed. Thus the spinor representation of complex states splits the states so that the positive energy part of the spinor with unprimed indices has negative helicity, while the positive energy part of the spinor with primed indices has positive helicity.

1.3. *Momentum and angular momentum*

In special relativistic dynamics any finite system possesses a total momentum p^a (a 4-vector) and a total angular momentum M^{ab} (skew tensor) dependent on the origin O. If $O \to \tilde{O}$, then $p^a \to \tilde{p}^a = p^a$ and $M^{ab} \to \tilde{M}^{ab} = M^{ab} - 2X^{[a}p^{b]}$ where X^a is the displacement $\overrightarrow{O\tilde{O}}$.

---

[*] For an empty region surrounding a source, this also holds globally if and only if a certain ten integrals vanish [14].

b) Spin $|s| \neq 0$. In this case all points on the null hyperplane $P^a X_a = k$ turn out to be on an equal footing. That is, one can find Poincaré transformations under which any two given points are dynamically equivalent. In this sense *the particle is not localised.* However, if we take two points $a, b \in K$, the necessary Poincaré transformation demonstrating the equivalence of a and b is not simply a translation, but a translation plus a specific null rotation (Trautman in [35]).

The null vector P^a corresponds to a spinor $\pi_{A'}$

$$P_a = \bar{\pi}_A \pi_{A'} \tag{1.32}$$

uniquely up to phase; $\pi_{A'} \to e^{i\theta} \pi_{A'}$ preserves P_a. $M^{ab} (= M^{[ab]})$ is represented by a symmetric spinor, $\mu^{AB} = \mu^{(AB)}$,

$$M^{ab} = \mu^{AB} \epsilon^{A'B'} + \epsilon^{AB} \bar{\mu}^{A'B'}.$$

The equation $S^a = sP^a$ takes the form

$$S_{DD'} = -i\bar{\pi}_D \pi^{A'} \bar{\mu}_{A'D'} + i\bar{\pi}^A \mu_{AD} \pi_{D'} = s\,\bar{\pi}_D \pi_{D'}. \tag{1.33}$$

If we transvect this with $\bar{\pi}^D$ we find $\mu_{AB} \bar{\pi}^A \bar{\pi}^B = 0$ which implies that

$$\mu_{AB} = i\omega_{(A} \bar{\pi}_{B)}$$

for some ω^A. Since any symmetric 2-index spinor is the symmetrised outer product of two spinors[6], the only new information concerning μ_{AB} is that one of the factors is $\bar{\pi}_A$. Thus

$$M^{AA'BB'} = i\bar{\pi}^{(A} \omega^{B)} \epsilon^{A'B'} - i\pi^{(A'} \bar{\omega}^{B')} \epsilon^{AB}. \tag{1.34}$$

We can now characterise the pair (P_a, M^{ab}) by the two spinors $(\omega^A, \pi_{A'})$ (but not uniquely, for the same pair is represented by $(e^{i\theta} \omega^A, e^{i\theta} \pi_{A'})$). This pair is a (representation of a) *twistor*[10] Z^α. We do not choose to *define* the twistor Z^α as the pair of spinors $(\omega^A, \pi_{A'})$ since under change of origin (and under conformal rescaling) the ω^A and $\pi_{A'}$ become transformed, whereas the twistor Z^α is supposed to remain unaffected. Thus we must think of $(\omega^A, \pi_{A'})$ simply as a *representation* of the twistor Z^α. In fact twistors have two stages of representation. The first, in terms of the above pair of spinors, is specified by a given origin and choice of conformal scale, i.e. of one of the conformally related flat metrics. (The spinor indices are here just abstract labels[1].) The second is in terms of the coordinates of these spinors with respect to some spinor frame. Such coordinates will be indicated by the presence of Hebrew indices.

If Z^α is represented by $(\omega^A, \pi_{A'})$ then we can take twistor components

$$Z^{\aleph} = (\omega^0, \omega^1, \pi_{0'}, \pi_{1'}). \tag{1.35}$$

We define a conjugate twistor $\bar{Z}_\alpha$ to have components

$$\bar{Z}_{\aleph} = (\bar{\pi}_0, \bar{\pi}_1, \bar{\omega}^{0'}, \bar{\omega}^{1'}). \tag{1.36}$$

Dropping the spinor frame, we have the representations $Z^\alpha \leftrightarrow (\omega^A, \pi_{A'})$, $\bar{Z}_\alpha \leftrightarrow (\bar{\pi}_A, \bar{\omega}^{A'})$ and so

$$Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^{A'} = 2s \tag{1.37}$$

---

[10] Note that as twistors will have well-defined translational transformations, they differ from Dirac spinors, see [5]. Quantities that are essentially twistors have been described by other authors [16].

We may define the spin vector

$$S_a := -\tfrac{1}{2}\eta_{abcd}P^b M^{cd} \tag{1.29}$$

Then

$$\widetilde{S}_a = S_a.$$

i) Assume $P_c P^c > 0$. Then the relativistic centre of mass of the system is defined to move on the worldline which is the locus of origins $\widetilde{\mathrm{O}}$ such that

$$\widetilde{P}_a \widetilde{M}^{ab} = 0. \tag{1.30}$$

Then, as regards its total momentum and angular momentum the system behaves as a single particle moving along this worldline with momentum $\widetilde{P}_a$ and intrinsic spin $\widetilde{M}^{ab}$. Equation (1.30) may be solved for X^a as

$$X^a = M^a{}_b P^b/(P_c P^c) + \lambda P^a$$

and this gives a unique timelike worldline. The intrinsic spin is

$$\widetilde{M}^{ab} = \eta^{abcd} S_c P_d/(P^e P_e).$$

ii) However we wish to consider zero rest mass (i.e. $P^c P_c = 0$) to be the more fundamental case. Then (1.30) has no solution unless

$$M_{ab}P^b = -P_a(P^b X_b).$$

Thus there is no solution unless

$$\begin{aligned} M^{ab}P_b \propto P^a &\Leftrightarrow P^{[a}M^{b]c}P_c = 0 \Leftrightarrow P^{[a}M^{bc]}P_c = 0 \\ &\Leftrightarrow P_{[c}S_{d]} = 0 \\ &\Leftrightarrow S_d = sP_d \end{aligned}$$

for some constant s, the helicity, whose modulus $|s|$ is the spin[*]. (This equation may also be deduced from other points of view about particles.) $X^b P_b = k$, k being a constant, is a null hyperplane K, so it appears that the "centre of mass line" has become a 3-dimensional region. We can say a little more by considering two cases separately:

a) Spin $|s| = 0$. Then $M^{ab} = 2A^{[a}P^{b]}$ where A^a is some vector, and the centre of mass line can be defined as

$$X^a = A^a + \lambda P^a,$$

the angular momentum about a point on this line being

$$\widetilde{M}^{ab} = 0.$$

Thus we may pick a specific generator of the hypersurface as the centre of mass line, and, as in the case where $P_a P^a > 0$ and $|s| = 0$, the system is completely characterised by this line.

---

[*] We shall choose units such that $h = 1$. Then for quantum systems $|s|$ takes half integer values.

using (1.33). Note that the Hermitian form used on the twistor Z^α in (1.37) has signature (++−−) and that positive ($s > 0$) and negative ($s < 0$) helicities are thus both possible.

When $s = 0$ the twistor is said to be null and represents a null worldline. When $s \neq 0$, the twistor represents a particle with intrinsic spin and there is a sense in which this means that the worldline is displaced into the complex. The particle ceases to be localised in M. The remaining sections will develop the twistor concept.

2. Twistors in flat space

2.1. *Basic ideas; conformal invariance*

In section 1.3 the concept of a twistor Z^α and its complex conjugate $\bar{Z}_\alpha$ were introduced, Z^α being represented by a pair of spinors $(\omega^A, \pi_{A'})$ which define the momentum and angular momentum of a massless particle by (1.32) and (1.34), $\bar{Z}_\alpha$ being correspondingly represented by $(\bar{\pi}_A, \bar{\omega}^{A'})$. The helicity of the particle is one-half the Hermitian norm $Z^\alpha \bar{Z}_\alpha$ of the twistor. Twistors also have a linear structure (i.e. $\lambda Z^\alpha \leftrightarrow (\lambda\omega^A, \lambda\pi_{A'})$; $\overset{1}{Z}{}^\alpha + \overset{2}{Z}{}^\alpha \leftrightarrow (\overset{1}{\omega}{}^A + \overset{2}{\omega}{}^A, \overset{1}{\pi}_{A'} + \overset{2}{\pi}_{A'})$) so we may expect the group of transformations which preserves these structures to have some significance. Since the signature of the form $Z^\alpha \bar{Z}_\alpha$ is (++−−), this group is U(2,2). But if we wish to retain the geometrical significance of the phase of a twistor in terms of a polarisation plane (i.e. the flag plane direction of $\pi_{A'}$) then we are led to consider the group SU(2,2) this being actually 4-1 homomorphic with the restricted conformal group. It turns out that twistors form a 4-1 representation space for the identity-connected component of the conformal group. The algebra of twistors is discussed in detail in [5]. Twistor space is 8-real-dimensional (4 complex dimensions). We may regard these dimensions as arising as follows; there is a five-dimensional set of null geodesics in M (consider the generators of the light cones with vertices at the points of any fixed spacelike surface) and on each geodesic one may give the momentum scaling (one parameter); the seventh dimension is the polarisation (phase of $\pi_{A'}$) and the eighth the intrinsic spin. The non-vanishing of intrinsic spin ($s \neq 0$) implies we do not have a uniquely-defined null geodesic, nor can we easily extend our interpretation to curved space-times (see section 3). Twistors are a sort of "square root" of the momentum and angular momentum in the same sense in which spinors are a "square root" of vectors.

In section 1.3 we said that a twistor was represented by a pair of spinors in a way dependent on choice of origin and conformal scale. We now need to know how the representation alters on change of origin and/or scale. Let us first consider the effect of a change of origin on the spinors which represent the twistor. When $\mathrm{O} \to \tilde{\mathrm{O}}$, we have $P_a \to \tilde{P}_a = P_a$; $M^{ab} \to \tilde{M}^{ab} = M^{ab} - 2X^{[a}P^{b]}$. If we further insist that the phase of $\pi_{A'}$ be unaltered on translation owing to its interpretation as a polarisation plane (flag plane) we find

$$\tilde{\pi}_{A'} = \pi_{A'}; \tilde{\omega}^A = \omega^A - \mathrm{i}\, X^{AA'}\pi_{A'}, \qquad (2.1)$$

i.e. given g_{ab}, $\tilde{\omega}^A$ is a function of position $(\tilde{\mathrm{O}})$ and may be regarded as a spinor *field*. (Actually it corresponds, in the case of *null* twistors, to a field of null directions of straight lines intercepting the worldline, see [5].) By (2.1)

$$\nabla^{A'(A}\tilde{\omega}^{B)} = 0. \qquad (2.2)$$

In fact the form (2.1) follows from (2.2). For (2.2) implies $\nabla^{A}_{A'}\tilde{\omega}^{B}$ is skew in AB; so also is $\nabla^{C}_{C'}\nabla^{A}_{A'}\tilde{\omega}^{B}$ and hence (since in flat space we may commute the derivative operators) this latter is skew in CB and therefore in CAB. Thus it is zero, so $\nabla^{A}_{A'}\tilde{\omega}^{B}$ is constant. If this constant is written $-\mathrm{i}\epsilon^{AB}\pi_{A'}$ (being skew in AB), the general solution of (2.2) is seen to be (2.1).

Since

$$\pi_{A'} = \tfrac{1}{2}\mathrm{i}\nabla_{AA'}\tilde{\omega}^{A} \tag{2.3}$$

the field $\tilde{\omega}^{A}$ completely defines the twistor. Moreover, by (1.19) etc., we see that (2.2) is invariant under a conformal rescaling with $\hat{\tilde{\omega}}^{B} = \tilde{\omega}^{B}$. Thus a spinor field $\tilde{\omega}^{B}$ satisfying (2.2) can be used as a conformally invariant definition of a twistor, which therefore tells us how a spinor representation of a twistor behaves under change of origin or scale. It should be noted that change of origin preserves $\pi_{A'}$ but alters ω^{A}, while conformal rescaling preserves ω^{A} but alters $\pi_{A'}$, the point being that, as viewed from the origin, conformal rescalings make infinity appear to be in a different place; the spinor $\pi_{A'}$ is associated with the vertex I_{0} of the null cone at infinity in the same way that ω_{A} is associated with the origin O.

If we define (dropping the ~ from here onward)

$$\mathcal{W}^{a} = \omega^{A}\bar{\omega}^{A'}$$

then

$$\nabla^{(a}\mathcal{W}^{b)} = \tfrac{1}{4}g^{ab}(\nabla_{c}\mathcal{W}^{c}) \tag{2.4}$$

so that $\mathcal{W}^{b}$ is a conformal Killing vector. However,

$$M^{ab} = \nabla^{[a}\mathcal{W}^{b]}$$

is not conformally invariant. Thus the angular momentum is not a conformal invariant, although $\mathcal{W}^{a}$ (or ω^{A}) is. In fact, from (1.20)

$$\hat{M}^{ab} = \Omega^{-2}(M^{ab} + 2\Upsilon^{[a}\mathcal{W}^{b]}).$$

This also follows from

$$\hat{\omega}^{A} = \omega^{A}; \qquad \hat{\pi}_{A'} = \pi_{A'} + \Upsilon_{AA'}\omega^{A} \tag{2.5}$$

and (1.34) which further implies that

$$s = \tfrac{1}{2}(\omega^{A}\bar{\pi}_{A} + \pi_{A'}\bar{\omega}^{A'}) \tag{2.6}$$

is conformally invariant. It is the conformal invariance of (2.2) and (2.6), together with linearity, which shows that twistors form a representation space (locally) for the conformal group.

Now let us consider the equation

$$\nabla^{(M}_{P'}\alpha^{AB\ldots L)} = 0 \tag{2.7}$$

which may be regarded as the many-index spinor equation generalising $\nabla^{(M}_{P'}\omega^{A)} = 0$. The equation is conformally invariant if its solution obeys

$$\hat{\alpha}^{AB\ldots L} = \alpha^{AB\ldots L}.$$

If we now form

$$\Psi_{AB\ldots D} = \alpha^{E\ldots L}\phi_{AB\ldots DE\ldots L} \tag{2.8}$$

where ϕ is a solution to (1.25) we find that $\Psi_{\ldots}$ satisfies (1.25) for a lower spin. In fact (2.7) has $\binom{3+n}{3}$ linearly independent solutions if α has n indices, as shown explicitly in [5], p. 362. Each solution in turn, for a given n, may be substituted in (2.8).

For example, in the case of linearised gravitation $\phi_{\ldots\ldots}$ we can form

$$\Psi_{AB} = \alpha^{CD}\phi_{ABCD}$$

which is a Maxwell field. We may ask what charge integrals this gives. There are 10 independent solutions for α^{CD}, so we will obtain 10 conserved (complex) quantities. These are in fact the energy, momentum and angular momentum [14]. (These quantities would be complex for a general solution of (1.25) but we get only 10 real quantities for a ϕ_{ABCD} derivable from a potential.) If the integrations are performed at infinity, these quantities give the Bondi-Sachs definition of mass [17], as applied to a general (shearing) retarded null hypersurface in Minkowski space, for linearised theory, so that it becomes clear that the "correction terms" which distinguish this mass measure from the Newman-Unti mass [18] are really necessary – even in linearised theory.

The equation

$$\nabla_{B'}^{(B}\omega^{A)} = 0 \tag{2.9}$$

which defines a twistor has 4 linearly independent solutions in M′. There is a difficulty at infinity because to form M we stick the past and future light cones together and the (one-superfix) twistors differ at those points by a factor i, essentially because the representation of the conformal group in twistor space is via a four-fold covering (cf. [19]). (For a many-index twistor, one must allow a factor i for each superfix and one factor –i for each suffix.) We could remove this difficulty by taking a fourfold covering of M but instead we simply adopt the rule of multiplying by i every time we complete a circuit passing through infinity. The problem is an illustration of the fact that twistors are like spinors in not being local geometric objects (for odd-indexed spinors are multiplied by –1 when they are rotated through 2π).

2.2. *Twistor space and Minkowski geometry*

A twistor with $2s = Z^\alpha \bar{Z}_\alpha = 0$ represents a null real straight line (i.e. the worldline of some particle of zero spin). If $s \neq 0$ there is no such real line, but there is in a certain sense a "complex line". Clearly when $s = 0$, Z^α and λZ^α $(\lambda \neq 0)$ represent the same line so that the most directly geometrically interpretable twistor space is the space N of equivalence classes $\{\lambda Z^\alpha\}$ when $s = 0$, $Z^\alpha \neq 0$, i.e.

$$\mathrm{N} = \{\{\lambda Z^\alpha : \lambda \neq 0, \lambda \in \mathbb{C}\} : Z^\alpha \bar{Z}_\alpha = 0, Z^\alpha \neq 0\}\ , \tag{2.10}$$

which represents the set of null lines in M. We shall therefore consider the space C of equivalence classes of twistors, defined like N but without the requirement $s = 0$ (fig. 3). This is complex projective three-space[11] $\mathbb{C}$P(3) which has three complex or six real dimensions. It is not just the com-

---

[11] For matters pertaining to projective geometry see [20].

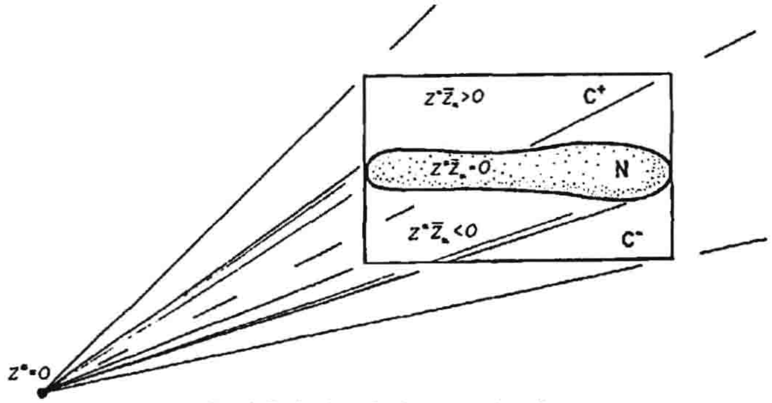

$Z^\alpha \bar{Z}_\alpha > 0$
C^+
$Z^\alpha \bar{Z}_\alpha = 0$
N
$Z^\alpha \bar{Z}_\alpha < 0$
C^-
$Z^\alpha = 0$

Fig. 3. Projection of twistor space into C.

plexification of N, which would have ten real dimensions. In fact even the complex points of C may be represented as real structures[12] (Robinson congruences) in M [5, 21, 22]. The conformal transformations of M correspond to projective point transformations[11] of C preserving N.

Let us now consider what a point in M corresponds to in C. We may define a point in M by the intersection of null lines. Suppose two lines are described by twistors $Z^\alpha \leftrightarrow (\omega^A, \pi_{A'})$, $Y^\alpha \leftrightarrow (\xi^A, \eta_{A'})$. Then the lines meet if there is a common solution to

$$\omega^A = \mathrm{i}p^{AA'}\pi_{A'}; \qquad \xi^A = \mathrm{i}p^{AA'}\eta_{A'}.$$

Formally this is

$$p^{AA'} = \frac{-\mathrm{i}}{\pi_{B'}\eta^{B'}}(\omega^A\eta^{A'} - \xi^A\pi^{A'}) \tag{2.11}$$

but of course the corresponding point need not be real, i.e. $p^{AA'}$ need not be "Hermitian". If the lines do meet (i.e. p^a real) then

$$\bar{\eta}_A\omega^A = \mathrm{i}\bar{\eta}_A p^{AA'}\pi_{A'} = \mathrm{i}\bar{\eta}_A \bar{p}^{AA'}\pi_{A'}\pi_{A'} = -\bar{\xi}^{A'}\pi_{A'}$$

i.e. $Z^\alpha \bar{Y}_\alpha = 0$.

Thus the necessary conditions for the two twistors to represent real intersecting lines are

$$Y^\alpha \bar{Y}_\alpha = 0; \qquad Z^\alpha \bar{Z}_\alpha = 0; \qquad Z^\alpha \bar{Y}_\alpha = 0. \tag{2.12}$$

These are also sufficient [5] if we interpret the condition appropriately when Y and Z are parallel[13], in which case they meet at infinity, i.e. lie in a null hyperplane [5]. This can be shown, assuming

---

[12] Note that the spinor field representation (section 2.1) represents twistors up to a factor ± i or ± 1, i.e. it includes proportionality [5].

[13] We shall use boldface kernel letters, e.g. **Z**, to represent the twistor up to proportionality, and the corresponding geometric structures in both C and M pictures, reserving indexed letters, e.g. Z^α, for the actual twistors, etc..

Y and Z to be non-parallel so that π^A and η^A are not proportional (and then taking a limit for the parallel case), by testing the Hermiticity of (2.11) by taking components with respect to π^A and η^A. The three conditions thus derived for Hermiticity are simply (2.12).

If (2.12) holds, then it is also satisfied if Y^α (or Z^α) is replaced by

$$X^\alpha = \mu Z^\alpha + \lambda Y^\alpha$$

for any complex numbers μ, λ. Thus the line X meets each of Y and Z and so belongs to the null cone through the point P with position vector p^a. This null cone can be used to represent P. Thus P is represented in N by the linear set $\mu X^\alpha + \lambda Y^\alpha$, i.e. by the (complex) line P joining points Z and Y (which has 2 real dimensions and topology S^2).

We may therefore represent this point P by the 2-index twistor

$$P^{\alpha\beta} = Z^\alpha Y^\beta - Y^\alpha Z^\beta \leftrightarrow \begin{pmatrix} \omega^A \xi^B - \xi^A \omega^B & \omega^A \eta_{B'} - \xi^A \pi_{B'} \\ \pi_{A'} \xi^B - \eta_{A'} \omega^B & \pi_{A'} \eta_{B'} - \eta_{A'} \pi_{B'} \end{pmatrix}$$

$$= \pi_{D'} \eta^{D'} \begin{pmatrix} -\frac{1}{2} \epsilon^{AB} p_{CC'} p^{CC'} & \mathrm{i} p^A{}_{B'} \\ -\mathrm{i} p_{A'}{}^B & \epsilon_{A'B'} \end{pmatrix} \tag{2.13}$$

where $p^a = p^{AA'}$ is the position vector of the point P. Thus the points of M correspond (up to proportionality) to simple skew 2-index twistors, i.e. twistors obeying

$$P^{\alpha\beta} = P^{[\alpha\beta]}; \qquad P^{[\alpha\beta}P^{\gamma]\delta} = 0 \text{ (i.e. } P^{[\alpha\beta}P^{\gamma\delta]} = 0).$$

We may define the dual twistor $P_{\alpha\beta}$ (which gives the geometrically dual description of the same line[11]) by

$$P_{\alpha\beta} = \tfrac{1}{2} P^{\rho\sigma} \epsilon_{\alpha\beta\rho\sigma}. \tag{2.14}$$

One may verify that p^a is real $\Leftrightarrow P_{\alpha\beta} = \overline{P}_{\alpha\beta}$ where $\overline{P}_{\alpha\beta}$ is the twistor complex conjugate of $P^{\alpha\beta}$. More generally, if p^a is complex, then its complex conjugate $\overline{p}^a$ corresponds to $\overline{P}^{\alpha\beta}$ in the same way that p^a corresponds to $P^{\alpha\beta}$. In fact the imaginary part of p^a is spacelike, timelike or null respectively according as P intersects N in a one real-dimensional set (a curve: topology S^1), in a point, or not at all. If null or timelike, the imaginary part of p^a is future pointing or past-pointing according as P lies in $C^- \cup N$ or $C^+ \cup N$.

Now we recall that we are working in compactified Minkowski space (fig. 1). Suppose P is in fact I, the vertex of the null cone at infinity. Null lines at infinity have $\pi_{A'} = 0 = \eta_{A'}$ and so the point I corresponds to the twistor

$$I^{\alpha\beta} \leftrightarrow \begin{pmatrix} \epsilon^{AB} & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.15}$$

This we shall call the 'infinity twistor'. Its dual is

$$I_{\alpha\beta} \leftrightarrow \begin{pmatrix} 0 & 0 \\ 0 & \epsilon^{A'B'} \end{pmatrix}.$$

We can normalise skew twistors by

$$P^{\alpha\beta}I_{\alpha\beta} = 2 \qquad (\text{i.e. } \pi_{A'}\eta^{A'} = 1). \tag{2.16}$$

This fails only if $\pi_{A'}$ and $\eta_{A'}$ are proportional, i.e. only if **P** is at infinity.

Suppose now that $P^{\alpha\beta}I_{\alpha\beta} = Q^{\alpha\beta}I_{\alpha\beta} = 2$. Then by direct calculation we obtain

$$P^{\alpha\beta}Q_{\alpha\beta} = -(p^a - q^a)(p_a - q_a) =: -(PQ)^2 \tag{2.17}$$

(or for non-normalised twistors

$$4P^{\alpha\beta}Q_{\alpha\beta}/(I_{\rho\sigma}P^{\rho\sigma})(Q^{\delta\gamma}I_{\delta\gamma}) = -(p^a - q^a)(p_a - q_a)).$$

This is clearly a Poincaré invariant quantity. In fact the subgroup of the conformal group which leaves $I_{\alpha\beta}$ invariant is just the Poincaré group. We can form a conformal invariant from the twistors of four points, namely

$$\Phi = (P^{\alpha\beta}Q_{\alpha\beta})(R^{\rho\sigma}S_{\rho\sigma})/(P^{\lambda\mu}S_{\lambda\mu})(R^{\tau\nu}Q_{\tau\nu}) = (PQ)^2(RS)^2/(PS)^2(RQ)^2$$

which defines a sort of "cross-ratio" for any four points in M.

To sum up, a general (complex projective) line in the (projective) twistor space C corresponds to a point in ℂM, the complexification of M; a line in N corresponds to a real point in M; a point in N corresponds to a null line in M. Starting from the space C we can reconstruct ℂM as the Klein representation[11] of lines in the complex three-dimensional projective space C, giving ℂM as a quadric fourfold in five-dimensional projective space. (In this case all dimensions are complex, so ℂM has eight real dimensions, cf. [23].)

2.3. *Solutions of the zero-rest-mass field equations*

The question we now discuss is how fields in M are represented in twistor space. We shall find that the general zero-rest-mass free fields can be remarkably concisely represented by holomorphic (complex analytic) functions $g(Z^\alpha)$ and $f(W_\alpha)$ on the twistor space and its dual[14], C*. But in order to make the correspondence we must take suitable contour integrals. Thus only the residues at the poles of f will be physically meaningful; consequently the subsequent formalism will be based on contour integration in C.

The solutions of the equations (1.25) can be represented by a set of quantities $\phi_r(\text{P}; o^A, \iota^B)$ where $r = 0, 1, \ldots n$; o^A, ι^B are a pair of basis spinors at the point P, and

$$\phi_r = \phi_{AB\ldots L}\,\underbrace{\iota^A\ldots\iota^D}_{r}\;\underbrace{o^E\ldots o^L}_{n-r}\;.$$

Now o_A and ι_B define null twistors through P, namely[14] U_α, V_β say, i.e. $U_\alpha \leftrightarrow (o_A, -\mathrm{i}p^{AA'}o_A)$, $V_\beta \leftrightarrow (\iota_B, -\mathrm{i}p^{BB'}\iota_B)$. Thus we have the quantities

$$\Phi_r(U_\alpha, V_\beta) = \phi_r(\text{P}; o^A, \iota^B), \qquad r = 0, \ldots n.$$

---

[14] There is no essential reason why we should not represent our geometric objects using the space of dual twistors, rather than the space of twistors; there is a quite free choice between these spaces. However, with the choices of convention we have made, *it turns out that spinor fields with unprimed indices, whose positive energy parts represent left-handed particles, must be represented by functions on the dual twistor space, while spinors with primed indices, whose positive energy parts represent right-handed particles (see section 1.2) will be represented by functions on the twistor space.*

If U_α and V_β are restricted to be null twistors with real intersection, Φ_r represent a zero-rest-mass field in M. Such a field may be regarded as defined on some *three*-parameter initial set (Cauchy hypersurface) and thence extended over the rest of space by the field equations. In twistor terms it would be economical if we could describe the field on M by some field on the (complex) 3-space C, or C*. So far it appears that we must define the field on *pairs* of points U, V in C*.

Let us take the point P and define a standard tensor and spinor reference frame (cf. (1.6)) such that

$$u = p^{00'} = \frac{p^0 + p^1}{\sqrt{2}}; \qquad \zeta = p^{01'} = \frac{p^2 + \mathrm{i}p^3}{\sqrt{2}}$$

$$\tilde{\zeta} = p^{10'} = \frac{p^2 - \mathrm{i}p^3}{\sqrt{2}}, \qquad v = p^{11'} = \frac{p^0 - p^1}{\sqrt{2}}.$$

$\bar{\zeta} = \tilde{\zeta}$, $u = \bar{u}$, $v = \bar{v}$ if and only if p^a is real. The field equations (1.25) become

$$\partial\phi_r/\partial\tilde{\zeta} = \partial\phi_{r+1}/\partial u; \qquad \partial\phi_r/\partial v = \partial\phi_{r+1}/\partial\zeta; \qquad r = 0, \ldots n-1. \tag{2.18}$$

These equations are automatically satisfied if

$$\phi_r = \frac{1}{2\pi\mathrm{i}}\oint_{\mathrm{K}} \lambda^r F(\lambda, u + \lambda\tilde{\zeta}, \zeta + \lambda v)\mathrm{d}\lambda \tag{2.19}$$

where F is a holomorphic (i.e. analytic or regular in the complex sense) function of three complex variables, the contour K being taken to surround the poles of F in a suitable way. The resulting fields will always be analytic in the real sense with respect to u, v, ζ, $\tilde{\zeta}$, but we may represent non-analytic fields as limits of analytic ones.

A real null vector at $p^a = (u, v, \zeta, \tilde{\zeta})$ has direction given by $\mathrm{d}u : \mathrm{d}v : \mathrm{d}\zeta : \mathrm{d}\tilde{\zeta}$ where

$$\mathrm{d}u + \lambda\mathrm{d}\tilde{\zeta} = 0 = \mathrm{d}\zeta + \lambda\mathrm{d}v$$

for some complex λ (possibly infinite). For the Minkowski metric is $2(\mathrm{d}u\,\mathrm{d}v - \mathrm{d}\zeta\,\mathrm{d}\tilde{\zeta})$ so that $\mathrm{d}u\,\mathrm{d}v = \mathrm{d}\zeta\,\mathrm{d}\tilde{\zeta}$ for a null direction. Thus $\mathrm{d}u : \mathrm{d}v : \mathrm{d}\zeta : \mathrm{d}\tilde{\zeta} \doteq \lambda\bar{\lambda} : 1 : -\lambda : -\bar{\lambda}$. The corresponding (null) twistor is $U_\alpha + \lambda V_\alpha = W_\alpha \leftrightarrow (\bar{\pi}_A, \bar{\omega}^{A'})$ where

$$\bar{\pi}_{\mathfrak{A}}\pi_{\mathfrak{A}'} \propto \begin{pmatrix} \mathrm{d}v & -\mathrm{d}\bar{\zeta} \\ -\mathrm{d}\zeta & \mathrm{d}u \end{pmatrix}$$

and $\lambda = \bar{\pi}_1/\bar{\pi}_0 = W_1/W_0$. Thence, as $\bar{\omega}^{A'} = -\mathrm{i}p^{AA'}\pi_{A'}$,

$$(W_2, W_3) = (\bar{\omega}^{0'}, \bar{\omega}^{1'}) = -\mathrm{i}(\bar{\pi}_0, \bar{\pi}_1)\begin{pmatrix} u & \zeta \\ \bar{\zeta} & v \end{pmatrix} = -\mathrm{i}\, W_0(u + \lambda\bar{\zeta}, \zeta + \lambda v).$$

Thus $(W_0, W_1, W_2, W_3) = W_0(1, \lambda, -\mathrm{i}(u + \lambda\bar{\zeta}), -\mathrm{i}(\zeta + \lambda v))$. If we therefore set

$$f(W_\alpha) = (W_0)^{-n-2}\, F(W_1/W_0, \mathrm{i}W_2/W_0, \mathrm{i}W_3/W_0)$$

then $f(W_\alpha)$ is homogeneous of degree $-n-2$ in W_α. (We can now check that this has the correct

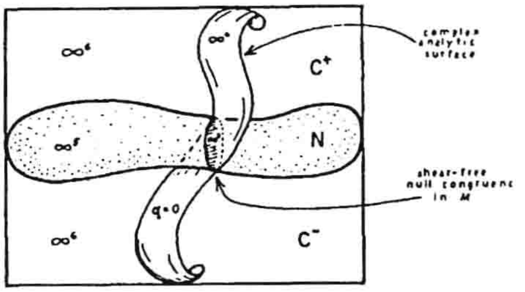


Fig. 4. The Kerr theorem.

transformation properties under rotation[15] for spin $\frac{1}{2}n$.) The final formula is

$$\Phi_r(U_\alpha, V_\beta) = \frac{1}{2\pi i} \oint_K \lambda^r f(U_\alpha + \lambda V_\alpha) d\lambda. \tag{2.20}$$

We may now generalise by taking *any* U_α, V_β (no longer necessarily null) thus defining complex fields on complex points $U_{[\alpha} V_{\beta]}$. It seems (although there is as yet no completely satisfactory theorem) that the set of such fields is extremely general. For a particular field it is clear that f is not unique since all the contour integrals remain the same under $f \to f + h$ where h is regular inside the contour. We may regard this as a sort of gauge invariance. This non-uniqueness of f would clearly lead to difficulties for any proposed explicit formula giving f in terms of $\phi_{A \dots L}$.

It is however easy to construct special types of solution for f. For example $\phi_{A \dots L}$ is called null if

$$\phi_{AB \dots L} = \alpha_A \alpha_B \dots \alpha_L$$

and such a field arises when the contour surrounds only a single simple pole [24]. (Note that a general symmetric spinor may be written as a symmetrised product of one-spinors[6] [2, 9].) More generally, the algebraically special fields

$$\phi_{AB \dots L} = \alpha_{(A} \alpha_B \beta_C \dots \lambda_{L)}$$

appear as integrals round contours surrounding a pole of order $\leqslant (n-1)$. E.g. to obtain type {2,2} (i.e. Petrov type D) linearized Weyl tensor fields we may require that f has two triple poles and that the contour separates one of them from the other. (Since here f has homogeneity degree -6, it follows that such an f is in fact the inverse cube of a quadratic form.)

If ϕ is algebraically special (e.g. null) there is associated with it a shearfree null congruence [22]. If

$$f(W_\alpha) = p(W_\alpha)/q(W_\alpha)$$

then $q(W_\alpha) = 0$ is a four (real) dimensional surface in a six dimensional space (C), and intersects the 5-dimensional surface N in a 3-dimensional set of points (fig. 4). This represents a 3-parameter null congruence in M. By a theorem of R.P. Kerr (unpublished, see [21]), this congruence must be shearfree. The theorem is that a congruence of null lines is shearfree if and only if it is representable in C as the intersection of N with a complex analytic surface S in C (or as a limiting case of such an intersection). It was partly this theorem that motivated the study of holomorphic functions in twistor space.

---

[15] See [21] for more details.

If we suppose $q = 0$ is a plane (i.e. $q(W_\alpha) = A^\alpha W_\alpha$) then we obtain by the above method a "linear" system of null lines in M (a Robinson congruence [5]), which we may consider to be a geometrical picture of the (complex) twistor A^α (which previously had no intuitively obvious picture associated with it). These "Robinson" congruences are largely what led to the name twistor, for they are shearfree, and twist with a handedness dependent on the sign of $A^\alpha \bar{A}_\alpha$.

If we consider a sourcefree spin $\frac{1}{2}n$ massless field in M (compactified Minkowski space), which has the correct peeling-off behaviour towards infinity [4], then the field will not match at infinity [19] unless we take a fourfold covering for odd n (twofold for $n \equiv 0 \bmod 4$). (This is reflected in the behaviour of the integrals introduced above since the homogeneity degree of $f(Z)$ is $-n-2$ and twistors are 4-valued, see section 2.1.) Rather than work with awkward covering spaces, however, we shall make the convention that a source-free field with the correct peeling-off properties is to be regarded as continuous across infinity if it has the right "Grgin discontinuity" at infinity (i.e. a general free wave of spin $\frac{1}{2}n$ should jump by a factor of i^{n+2} [19, 21]).

Consider then fields with the correct peeling-off and Grgin behaviour (which momentum eigenstates, for example, do not have). These may be (uniquely) split into positive and negative energy fields (cf. [25]). A process equivalent to Grgin's harmonic analysis technique [19] applied to the positive energy fields is the following. Instead of $\bar{Z}_0 = \bar{Z}^2$ etc., let us take twistor coordinates so that we get the more natural-looking $\bar{Z}_\aleph = (\bar{Z}^0, \bar{Z}^1, -\bar{Z}^2, -\bar{Z}^3)$, the Hermitian form $Z^\alpha \bar{Z}_\alpha$, of signature (++−−), being now diagonalised. The orthonormal basis $\{E_\alpha\}$ then has two vectors of positive and two of negative length. These points give us four planes (fig. 5) and the simplest possible function of positive frequency has as its singular region just the planes shaded in fig. 5 (cf. [21]). A general function for spin $\frac{1}{2}n$ fields of positive frequency is

$$f(\bar{Z}_\alpha) = \sum_{a_0 a_1 a_2 a_3} \frac{(\bar{Z}_0)^{a_0} (\bar{Z}_1)^{a_1}}{(\bar{Z}_2)^{a_2+1} (\bar{Z}_3)^{a_3+1}} f_{a_0 a_1 a_2 a_3} \tag{2.21}$$

where $f_{a_0 a_1 a_2 a_3}$ is a constant and a_0, a_1, a_2, a_3 are non negative integers satisfying $a_0 + a_1 + n = a_2 + a_3$. If S is the set of singularities of this function then assuming suitable convergence $S \cap C^{-*}$ is disconnected in two pieces, and so will yield a positive frequency field [21]. The individual terms in (2.21) will in fact form an orthogonal basis according to the scalar product of section 3.3.

For a further discussion of the material of this section see [21, 24].

2.4. *Quantisation*

We start out by considering how to connect the spin s of relativistic dynamics, which appeared in the classical twistor picture of angular momentum discussed above (section 1.3) with the spin s of the zero-rest-mass fields just considered.

The momentum of a particle with zero-spin was described by $\pi_{A'} (\bar{\pi}_A \pi_{A'} = p_a)$ while the position of the centre of mass is then determined by $\omega^A = iX^{AA'} \pi_{A'}$. As

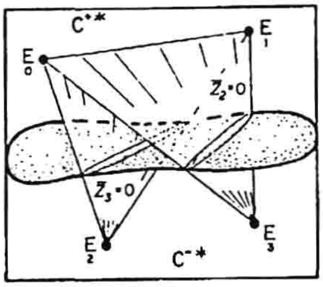


Fig. 5.

$$Z^\alpha \leftrightarrow (\omega^A, \pi_{A'}) \qquad \bar{Z}_\alpha \leftrightarrow (\bar{\pi}_{A'}, \bar{\omega}^{A'})$$

we find that

$$\begin{aligned} \mathrm{i}Z^\alpha \mathrm{d}\bar{Z}_\alpha &\leftrightarrow \mathrm{i}\omega^A \,\mathrm{d}\bar{\pi}_A + \mathrm{i}\pi_{A'}\mathrm{d}\bar{\omega}^{A'} \\ &= -X^{AA'}\pi_{A'}\mathrm{d}\bar{\pi}_A + \pi_{A'}\mathrm{d}(X^{AA'}\bar{\pi}_A) \\ &= -X^{AA'}\pi_{A'}\mathrm{d}\bar{\pi}_A + \pi_{A'}(\mathrm{d}X^{AA})\bar{\pi}_A + \pi_{A'}X^{AA'}\mathrm{d}\bar{\pi}_A \\ &= \pi_{A'}\bar{\pi}_A \,\mathrm{d}X^{AA'} = P_a \mathrm{d}X^a \end{aligned} \tag{2.22}$$

if $X^{AA'}$ is real. Thus, taking the exterior derivative,

$$\mathrm{i}\mathrm{d}Z^\alpha \wedge \mathrm{d}\bar{Z}_\alpha = \mathrm{d}P_a \wedge \mathrm{d}X^a \tag{2.23}$$

and the right-hand side is just the two-form preserved under canonical transformations, i.e. by Hamiltonian equations. (For a fuller account of this correspondence see [26].) This suggests that we should regard $-\mathrm{i}Z^\alpha$, $\bar{Z}_\alpha$ as canonically conjugate variables. Thus in the passage to a quantum theory we would expect $-\mathrm{i}Z^\alpha$, $\bar{Z}_\alpha$ to become canonically conjugate operators (with $\bar{Z}_\alpha \propto \partial/\partial Z^\alpha$, etc.).

In the operator form

$$\begin{aligned} &\mathbf{P}_a = \mathrm{i}\,\partial/\partial x^a \qquad (\text{and } \mathbf{X}^a = -\mathrm{i}\,\partial/\partial P_a) \\ &\mathbf{P}_a \mathbf{X}^b - \mathbf{X}^b \mathbf{P}_a = \mathrm{i}\,\delta_a^b, \end{aligned} \tag{2.24}$$

units being chosen so that $\hbar = 1$. Thus we shall want

$$\mathrm{Z}^\alpha = \partial/\partial \bar{\mathrm{Z}}_\alpha \qquad (\bar{\mathrm{Z}}_\alpha = -\partial/\partial \mathrm{Z}^\alpha)$$

and

$$\mathrm{Z}^\alpha \bar{\mathrm{Z}}_\beta - \bar{\mathrm{Z}}_\beta \mathrm{Z}^\alpha = \delta_\beta^\alpha, \tag{2.25}$$

where these operators are taken to act on functions $f(\bar{Z}_\alpha)$. Now in the method of section 2.3, ϕ is essentially given by $f(\bar{Z}_\alpha)$, and it is clear from taking complex conjugates that solutions of (1.28) are similarly described by a function $g(Z^\alpha)$. Now

$$Z^\alpha f(\bar{Z}) = \frac{\partial}{\partial \bar{Z}_\alpha} f(\bar{Z}); \qquad \bar{Z}_\alpha f(\bar{Z}) = \bar{Z}_\alpha f(\bar{Z})$$

$$Z^\alpha g(Z) = Z^\alpha g(Z); \qquad \bar{Z}_\alpha g(Z) = -\frac{\partial}{\partial Z^\alpha} g(Z). \tag{2.26}$$

Previously we had $Z^\alpha \bar{Z}_\alpha = 2s$, where $S^\alpha = sP^\alpha$, s being the spin parallel to the direction of motion. So consider the operator S defined by

$$4S := Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha = 2(\bar{Z}_\alpha Z^\alpha + 2) = 2(Z^\alpha \bar{Z}_\alpha - 2). \tag{2.27}$$

Then

$$Sg(Z^\alpha) = \tfrac{1}{2}((n+2) - 2)g(Z^\alpha) = sg(Z^\alpha)$$

for g is homogeneous of degree $(-n-2)$ and $2s = n$ whereas $Z^\alpha \partial g(Z)/\partial Z^\alpha$ gives $kg(Z^\alpha)$ where k is the homogeneity degree. (One may, incidentally, say that the fact that $\delta^\alpha_\alpha = 4$ in twistor space, i.e. its 4-dimensionality, is related to the need for the degree $(-n-2)$ in the definition of f.) We also find $Sf(\bar{Z}_\alpha) = sf(\bar{Z}_\alpha)$ if $n = -2s$, so that the twistor fields corresponding to spinors with primed indices are of opposite helicity, as we expect. The fact that the spin is half-integral is a consequence of the one-valuedness of f.

We may inquire what is the effect of Z^α, $\bar{Z}_\alpha$ when acting on the fields $\phi_{\ldots}$. Consider

$$f(W_\alpha) \to (Q^\alpha W_\alpha) f(W_\alpha), \tag{2.28}$$

which is the result of $Q^\alpha \bar{Z}_\alpha$. If $Q^\alpha \leftrightarrow (Q^A, Q_{A'})$, eq. (2.28) corresponds to

$$\phi_{AB\ldots L} \to \widetilde{Q}^A \phi_{AB\ldots L} = \psi_{B\ldots L}, \tag{2.29}$$

where $\widetilde{Q}^A = Q^A - \mathrm{i}X^{AA'}Q_{A'}$, and $\psi_{B\ldots L}$ satisfies the zero-rest-mass field equations for spin $(n-1)$. Similarly, if $R_\alpha \leftrightarrow (R_A, R^{A'})$, the operator $R_\alpha Z^\alpha$ acts so that

$$f(W_\alpha) \to R_\beta \frac{\partial}{\partial W_\beta} f(W_\alpha); \tag{2.30}$$

$\phi_{AB\ldots L} \to \tfrac{1}{2}\mathrm{i}(n+1)\phi_{(AB\ldots L}\nabla_{M)M'}\widetilde{R}^{M'} + \mathrm{i}\widetilde{R}^{M'}\nabla_{M'M}\phi_{AB\ldots L} = \chi_{AB\ldots M}$, where $\chi_{AB\ldots M}$ is a solution of the zero rest mass field equations for spin $(n+1)$. Thus $\bar{Z}_\alpha$ raises, and Z^α lowers, the helicity by one-half.

3. Twistors in curved space

There are three approaches to the problem of generalising the twistor formalism to curved space-time. These lead to three somewhat different twistor concepts, which will be referred to as global twistors, local twistors and asymptotic twistors.

Global twistors are the most logical generalisation of ordinary flat-space twistors to a curved space-time. However, they suffer from two serious shortcomings, namely that they have only a rather weak structure (merely symplectic, rather than linear and/or complex analytic) and that only the null twistors can be precisely defined in geometrical terms, the existence of the non-null twistors being simply postulated.

Local twistors, on the other hand, have a well-defined existence (whether null or non-null) and a linear and complex analytic structure. But they are necessarily defined relative to points in the space-time so they cannot in themselves be regarded as a satisfactory generalisation of the flat-space twistors, adequate to form a basis for a formalism in which space-time points are to be regarded as derived objects. Local twistors give rise to a conformally invariant calculus on a space-time manifold. This may have some utility as such; but the main value of the local twistor concept lies in its use in the definition of asymptotic twistors.

The asymptotic twistor concept is one which applies only to a space-time which is asymptotically flat. However this is the situation appropriate to an S-matrix theory of gravitation and consequently has great relevance for the twistor quantisation programme. The space of asymptotic twistors has a complex analytic structure with a (non-linear) Hermitian scalar product defined – giving rise to a (pseudo-) Kählerian (and hence also a symplectic) structure. Asymptotic twistors, which were developed after the lectures on which these notes are based were given, are discussed in section 5. A brief discription of local twistor theory will be given here and global twistor theory and its relation to gravitational scattering will be discussed.

3.1 *Local twistors*

In this approach we define a twistor space at each point of space-time[16]. This twistor space may be thought of as the direct sum of a spin-space and a conjugate spin space. However the exact way in which the twistor space splits up as a direct sum depends on the choice of conformal scaling. More explicitly, a local twistor Z^α at a point p can be represented, with respect to the metric g_{ab}, by a pair of spinors $(\omega^A, \pi_{A'})$ at p. Under a conformal rescaling (cf. section 2.1) we will have (cf. (2.5), (1.19a))

$$\hat{g}_{ab} = \Omega^2 g_{ab}; \qquad \hat{\omega}^A = \omega^A; \qquad \hat{\pi}_{A'} = \pi_{A'} + i\Upsilon_{AA'}\omega^A. \tag{3.1}$$

This is consistent with the behaviour already encountered in flat space-time since in that case we have

$$\hat{\nabla}_{AA'}\hat{\omega}^B = \hat{\nabla}_{AA'}\omega^B = \nabla_{AA'}\omega^B + \epsilon_A{}^B\Upsilon_{CA'}\omega^C \tag{3.2}$$

whence

$$i\hat{\epsilon}_A{}^B\hat{\pi}_{A'} = i\epsilon_A{}^B\pi_{A'} - \Upsilon_{CA'}\omega^C\epsilon_A{}^B \tag{3.3}$$

follows from

$$\nabla_{AA'}\omega^B = -i\epsilon_A{}^B\pi_{A'}. \tag{3.4}$$

However this last equation can only be maintained in conformally-flat space-time. Nevertheless it supplies motivation for (3.1) which will be retained for local twistors in curved space-time.

The calculation of covariant derivatives of local twistors is most easily accomplished by introducing projection and injection operators from the twistor space to the two spin-spaces which represent it. Thus we define[17]

$$e^A_\alpha, \; e_{\alpha A'}, \; e^\alpha_A, \; e^{\alpha A'}$$

[16] I.e. for one-superfix twistors we form a fibre bundle over space-time with fibre the (8-dimensional) twistor space.

[17] Here we follow unpublished work of K. Dighton, which is partially based on a suggestion of A. Qadir.

such that $Z^\alpha e^A_\alpha = \omega^A$, $Z^\alpha e_{\alpha A'} = \pi_{A'}$ etc. and

$$e^A_\alpha e^\alpha_B = \epsilon_B{}^A, \qquad e^A_\alpha e^{\alpha A'} = 0, \qquad e_{\alpha A'} e^{\alpha B'} = \epsilon_{A'}{}^{B'}, \qquad e_{\alpha A'} e^\alpha_A = 0,$$
$$e^A_\alpha e^\beta_A + e_{\alpha A'} e^{\beta A'} = \delta^\beta_\alpha, \qquad \overline{e^A_\alpha} = e^{\alpha A'}, \qquad \overline{e^\alpha_A} = e_{\alpha A'}. \tag{3.5}$$

Under conformal rescaling we will have $Z^\alpha \to \hat{Z}^\alpha = Z^\alpha$ while ω_A, $\pi^{A'}$ transform by (3.1). Thus we see that

$$\hat{e}^A_\alpha = e^A_\alpha; \qquad \hat{e}_{\alpha A'} = e_{\alpha A'} + \mathrm{i}\Upsilon_{AA'} e^A_\alpha,$$
$$\hat{e}^{\alpha A'} = e^{\alpha A'}; \qquad \hat{e}^\alpha_A = e^\alpha_A - \mathrm{i}\Upsilon_{AA'} e^{\alpha A'}. \tag{3.6}$$

We now have to decide how the projection operators vary as we pass from one point to another. We find we require (cf. (1.21))

$$\nabla_{RS'} e^{\alpha A'} = \mathrm{i}\epsilon_{S'}{}^{A'} e^\alpha_R, \qquad \nabla_{RS'} e^A_\alpha = -\mathrm{i}\epsilon_R{}^A e_{\alpha S'},$$
$$\nabla_{RS'} e^\alpha_A = \mathrm{i} P_{RAS'A'} e^{\alpha A'}, \; \nabla_{RS'} e_{\alpha A'} = -\mathrm{i} P_{RAS'A'} e^A_\alpha. \tag{3.7}$$

We have $\omega^A = e^A_\alpha Z^\alpha$, $\pi_{A'} = e_{\alpha A'} Z^\alpha$ so that

$$\nabla^\sigma_\rho Z^\alpha = e^{\sigma S'} e^R_\rho \; \nabla_{RS'}(e^\alpha_A \omega^A + e^{\alpha A'} \pi_{A'})$$
$$= e^R_\rho e^{\sigma S'} \{ e^\alpha_A (\nabla_{RS'} \omega^A + \mathrm{i}\epsilon_R{}^A \pi_{S'}) + e^{\alpha A'} (\nabla_{RS'} \pi_{A'} + \mathrm{i} P_{RBS'A'} \omega^B) \}. \tag{3.8}$$

These forms (3.7), (3.8) are required in order to give a conformally invariant twistor derivative as can be checked using (3.1) and (3.6), and because in flat space, constant local twistors (i.e. those annihilated by $\nabla_{RS'}$) will now correspond to our former global flat-space twistors[18]. When referred to a basis (3.8) has, in all, 64 components, 48 of them being zero. ∇^σ_ρ satisfies the usual requirements of a derivative (linearity and the Leibniz rule) and it commutes with complex conjugation and contraction.

We may now consider

$$\nabla^\sigma_\rho \nabla^\mu_\tau - \nabla^\mu_\tau \nabla^\sigma_\rho = [\nabla^\sigma_\rho, \nabla^\mu_\tau].$$

Acting on a scalar function ϕ, this gives us

$$[\nabla^\mu_\lambda, \nabla^\sigma_\rho] \phi = \mathrm{i}(\delta^\sigma_\lambda \nabla^\mu_\rho - \delta^\mu_\rho \nabla^\sigma_\lambda)\phi = T^{\mu\sigma\alpha}_{\lambda\rho\beta} \nabla^\beta_\alpha \phi \tag{3.9}$$

where

$$T^{\mu\sigma\alpha}_{\lambda\rho\beta} = \mathrm{i}(\delta^\sigma_\lambda \delta^\alpha_\rho \delta^\mu_\beta - \delta^\mu_\rho \delta^\sigma_\beta \delta^\alpha_\lambda).$$

Then

$$K^{\mu\sigma\alpha}_{\lambda\rho\beta} Z^\beta = \mathrm{i}\, \{ [\nabla^\mu_\lambda, \nabla^\sigma_\rho] - T^{\mu\sigma\gamma}_{\lambda\rho\beta} \nabla^\beta_\gamma \} \, Z^\alpha \tag{3.10}$$

---

[18] Eq. (3.8) shows that a constant local twistor in flat space has constant $\pi_{A'}$, and ω^A satisfying (3.4), i.e. we retrieve the spinor field representation of a twistor, and so the correspondence just referred to is that the local twistor at a point P has the same representation as the global twistor has when referred to P as origin.

where

$$K^{\mu\sigma\alpha}_{\lambda\rho\beta} = e^{L}_{\lambda} e^{\mu M'} e^{R}_{\rho} e^{\sigma S'} [e^{B}_{\beta} \{e^{\alpha}_{A} \mathrm{i}\epsilon_{M'S'} \Psi^{A}{}_{BLR} - e^{\alpha A'} (\epsilon_{RL} \nabla^{B'}_{B} \overline{\Psi}_{B'M'S'A'} - \epsilon_{M'S'} \nabla^{A}_{A'} \Psi_{ARLB})\} +$$
$$+ e_{\beta B'} e^{\alpha A'} \mathrm{i}\epsilon_{RL} \overline{\Psi}^{B'}{}_{A'M'S'}]. \qquad (3.11)$$

These define a torsion twistor $T^{\mu\sigma\alpha}_{\lambda\rho\beta}$ and a curvature twistor $K^{\mu\sigma\alpha}_{\lambda\rho\beta}$. The spinor components of $K^{\mu\sigma\alpha}_{\lambda\rho\beta}$ given by (3.11) involve Ψ_{ABCD} and $\nabla^{A}_{P'} \Psi_{ABCD}$.

Note that, by the conformal transformation rules for twistors (section 2.1), any local twistor $Q_{\alpha\beta}{}^{\gamma}$ for example, defines a conformally invariant spinor $Q^{A'B'G}$. If this spinor vanishes, then each of $Q_{A}{}^{B'G}$, $Q^{A'}{}_{B}{}^{G}$, $Q^{A'B'}{}_{G'}$ is conformally invariant. If $Q_{A}{}^{B'G} = 0$ and $Q^{A'}{}_{B}{}^{G} = 0$ then $Q_{AB}{}^{G}$ is conformally invariant, etc. etc.. Thus any non-vanishing local twistor defines at least one non-vanishing conformally invariant spinor. In particular we can apply this to $K^{\alpha\beta\gamma}_{\kappa\mu\nu}$ (to obtain $\epsilon_{B'G'} \Psi^{A}{}_{LMN}$) or to derivatives $\nabla^{\sigma}_{\rho} K^{\alpha\beta\gamma}_{\lambda\mu\nu}$, $\nabla^{\tau}_{\xi} \nabla^{\rho}_{\sigma} K^{\alpha\beta\gamma}_{\lambda\mu\nu}$ etc., or to such derivatives to which symmetry operations have been applied. (K. Dighton has shown how to obtain the Bach tensor and other conformally invariant tensors in this way.)

3.2. *Global twistors*

Consider a null geodesic Z with a parallelly propagated spinor $\pi_{A'}$ defined along it whose flagpole direction is tangent to Z. By analogy with the situation in flat space we may reasonably identify such a structure as a null (global) twistor in curved space. This description is conformally invariant (cf. (1.19a)). (We do not give a spinor representation of a twistor relative to each point[19], nor can we use our former description of a non-null twistor.) Null twistors form a seven-dimensional manifold N (5 dimensions for the set of null geodesics, 2 for the set of spinors $\pi_{A'}$). We shall consider N to be embedded in an abstract 8-dimensional manifold C, the points of C–N representing, formally, the non-null twistors. This is done because the structure of N is most easily described as that induced from the embedding of N in C, the structure of C being describable in simple terms. No precise geometrical definition of the elements of C–N will (can?) be given.

The space C will have a symplectic structure. Symplectic structures are only possible in even-dimensional spaces, and symplectic manifolds of the same dimension are locally congruent. The symplectic structure of C induces on N a structure which has geometric significance in the space-time. This structure on N expresses relations between neighbouring points of N, these relations representing geometrical connections between the null geodesics (and $\pi_{A'}$ spinors) that the points of N represent. Such geometrical connections must refer to properties of null geodesics. For example, the fact that a congruence of null geodesics has vanishing *rotation*, i.e. is (null) hypersurface forming, is such a property and it turns out that this property is simply describable in terms of the symplectic structure of C. The *shear* of a null congruence, on the other hand, is something which, in a general curved space-time, can be defined only in relation to points on the null geodesics. A null congruence which is shear-free at one point will, in the presence of conformal curvature, generally be shearing at other points. Recall that the Kerr theorem established a close connection between the shear-free condition for null congruences in flat space-time and the complex analytic structure of the C-picture. The fact that the concept "shear-free" cannot, in general curved space-times, be applied to null geodesics in their entirety, strongly indicates that C cannot generally be given a geometrically meaningful complex analytic structure.

---

[19] One might, for example, try to define $(\omega^{A}, \pi_{A'})$ at a point O by taking the intersection P of the null cone at O with the null geodesic (twistor) and parallelly propagating the π-spinor along the null geodesic PO or else using local twistor transport along PO. This does not appear to agree with the type of structure we shall require the twistor space to have.

Let us investigate the structure of C for a particular type of curved space-time $\mathcal{M}$, namely one which possesses two regions M_1 and M_2 of flat space-time separated by a curved region of $\mathcal{M}$, through which null geodesics can pass from M_1 to M_2. This will enable us to examine the structure of C in relation to the structure we have previously obtained for flat space-time. By involving two flat metrics we shall be able to isolate the structure of C as that which is common to the structures induced by each of M_1 and M_2. Now in each of M_1 and M_2 we can represent twistors in terms of pairs of spinors and hence in terms of four complex components Z^0, Z^1, Z^2, Z^3 (subject to $Z^{\aleph}\bar{Z}_{\aleph} = 0$). The expressions $Z^{\aleph}\mathrm{d}\bar{Z}_{\aleph}$ and $\mathrm{d}Z^{\aleph} \wedge \mathrm{d}\bar{Z}_{\aleph}$ define forms[20] on N which as we shall show are the same whether the coordinates $Z^{\aleph}$ are defined in M_1 or in M_2. Each of the forms $\phi = \mathrm{i}Z^{\aleph}\mathrm{d}\bar{Z}_{\aleph}$ and $\omega = \mathrm{d}\phi = \mathrm{i}\mathrm{d}Z^{\aleph} \wedge \mathrm{d}\bar{Z}_{\aleph}$ defines structure of geometrical significance in $\mathcal{M}$. It turns out, in fact, that ϕ measures time separation between neighbouring geodesics, while ω measures rotation, see also [26].

Let us consider two examples, both of flat spaces M_1, M_2 joined across a null hypersurface K, the (degenerate) metric of K being the same whether induced by M_1 or M_2. The curvature resides entirely within K, having the form of a δ-function on K.

A. Take two flat spaces

$$\begin{aligned} &M_1\colon\quad \mathrm{d}s^2 = 2(\mathrm{d}u\,\mathrm{d}v - \mathrm{d}\zeta\,\mathrm{d}\bar{\zeta}), && v \leqslant 0 \\ &M_2\colon\quad \mathrm{d}s^2 = 2(\mathrm{d}u^*\,\mathrm{d}v^* - \mathrm{d}\zeta^*\,\mathrm{d}\bar{\zeta}^*), && v \geqslant 0 \end{aligned} \tag{3.12}$$

joined on the null hyperplane $v = 0 = v^*$ where $\zeta^* = \zeta$; $u^* = u - q(\zeta\bar{\zeta})$. This has a δ-function in curvature on the join (rather as the surface of a cylinder of finite extent has, at the join of the end and the side – both of which are flat)[21]. The Ricci curvature is (essentially) $\delta(v)\partial^2 q/\partial\zeta\partial\bar{\zeta}$, while the conformal curvature is (essentially) $\delta(v)\partial^2 q/\partial\zeta^2$, $\delta(v)\partial^2 q/\partial\bar{\zeta}^2$. Einstein's empty space field equations yield

$$\partial^2 q/\partial\zeta\partial\bar{\zeta} = 0, \qquad \text{whence } q = r(\zeta) + \bar{r}(\bar{\zeta}), \tag{3.13}$$

r being a holomorphic (i.e. complex analytic) function.

B. Similarly join flat spaces

$$\begin{aligned} &M_1\colon\quad \mathrm{d}s^2 = \mathrm{d}u\,\mathrm{d}v - u^2\mathrm{d}\zeta\,\mathrm{d}\bar{\zeta}, && v \leqslant 0 \\ &M_2\colon\quad \mathrm{d}s^2 = \mathrm{d}u^*\mathrm{d}v^* - u^{*2}\mathrm{d}\zeta^*\mathrm{d}\bar{\zeta}^*, && v \geqslant 0 \end{aligned} \tag{3.14}$$

along $v = 0$ (a null cone) with $\zeta^* = f(\zeta)$ (f being a holomorphic function); $u^* = u/|f'(\zeta)|$. It turns out that this automatically satisfies Einstein's vacuum field equations. (For a fuller discussion of this case see [29].)

In these examples the passage of a null geodesic through $v = 0$ is determined by the condition that it is orthogonal to the same vectors within $v = 0$ on each side of the join. (The behaviour can also be found by considering an appropriate limit of C^{∞} spaces.) This tells us how a twistor is affected by an impulsive wave. In both cases the null geodesic is scattered in a way that can be formulated in Hamiltonian terms.

---

[20] For an introduction to Cartan forms see e.g. [27].

[21] This space is in fact a limiting case of the plane-fronted waves [28], see [21], while (3.14) is a limiting case of the Robinson-Trautman waves [31].

Let us see this explicitly in terms of example A. A twistor Z^α representing a null line, with coordinates as in section 2.3, has

$$Z^0:Z^1:Z^2:Z^3 = -iu\,d\zeta + i\zeta\,du : -i\bar{\zeta}\,d\zeta + iv\,du : -d\zeta : du$$

$$= -iu\,dv + i\zeta\,d\bar{\zeta} : -i\bar{\zeta}\,dv + iv\,d\bar{\zeta} : -dv : d\bar{\zeta}.$$

Thus it satisfies

$$-Z^3 d\zeta = Z^2 du, \qquad -Z^3 dv = Z^2 d\bar{\zeta}, \tag{3.15a}$$

$$Z^1 = i\bar{\zeta}Z^2, \qquad Z^0 = i\zeta Z^3 + iuZ^2 \tag{3.15b}$$

(since we are considering a point on K where $v = 0$). The starred version of (3.15b) also holds. Thus

$$Z^{*1} = i\bar{\zeta}Z^{*2}, \qquad Z^{*0} = i\zeta Z^{*3} + i(u - q)Z^{*2}. \tag{3.16}$$

In order to write the remainder of the starred version of (3.15) in terms of du, dv, $d\zeta$ we need to use the fact that Z^* and Z are orthogonal to the same vectors in K at the point $Z \cap K$. Denoting a direction at $Z \cap K$ by $\delta u : \delta v : \delta\zeta$ in the u, v, ζ system we have $\delta v = 0$ if the direction lies in K. For the direction to be orthogonal to that of Z we require

$$\delta u\, dv + 0\, du = \delta\zeta\, d\bar{\zeta} + \delta\bar{\zeta}\, d\zeta \tag{3.17}$$

whence, from (3.15)

$$\delta u = -\delta\zeta Z^3/Z^2 - \delta\bar{\zeta}\bar{Z}^3/\bar{Z}^2. \tag{3.18}$$

The starred version of this gives, from (3.12),

$$\delta u - \frac{\partial q}{\partial \zeta}\delta\zeta - \frac{\partial q}{\partial \bar{\zeta}}\delta\bar{\zeta} = -\delta\zeta Z^{*3}/Z^{*2} - \delta\bar{\zeta}\overline{Z^{*3}}/\overline{Z^{*2}}. \tag{3.19}$$

Equations (3.18) and (3.19) must represent identical conditions on $\delta u : \delta\zeta : \delta\bar{\zeta}$ since they must give the same 2-plane element. Hence

$$Z^3 : Z^2 = Z^{*3} - Z^{*2}\frac{\partial q}{\partial \zeta} : Z^{*2}. \tag{3.20}$$

Equations (3.15), (3.16) and (3.20) define the ratios of the $Z^{*\alpha}$ components in terms of the ratios of the Z^α components, by elimination of ζ and u. With the most convenient choice of scale factor we can set

$$Z^{*3} = Z^3 + Z^2\,\partial q/\partial\zeta; \qquad Z^{*2} = Z^2$$

$$Z^{*1} = Z^1, \qquad Z^{*0} = Z^0 - iZ^2(q - \zeta\,\partial q/\partial\zeta) \tag{3.21}$$

where $\zeta = i\bar{Z}^1/\bar{Z}^2$. Setting

$$H(Z^\alpha, \bar{Z}_\alpha) \equiv |Z^2|^2 q \tag{3.22}$$

we can write (3.21) comprehensively as

$$Z^{*\alpha} = Z^\alpha - \mathrm{i}\,\partial H/\partial \bar{Z}_\alpha. \tag{3.23}$$

The same formula, with H real and homogeneous of degree one separately in Z^α and in $\bar{Z}_\alpha$, is also valid for case B, though H now depends on $f(\zeta)$, rather than q. In the infinitesimal change case we find

$$\delta Z^\alpha = Z^{*\alpha} - Z^\alpha = -\mathrm{i}\,\partial H/\partial \bar{Z}_\alpha; \qquad \delta \bar{Z}_\alpha = \mathrm{i}\,\partial H/\partial Z^\alpha \tag{3.24}$$

which are equations of the Hamiltonian type and so preserve the symplectic structure [21, 26]. In fact $Z^\alpha \bar{Z}_\alpha$, $\phi = \mathrm{i}Z^\alpha \mathrm{d}\bar{Z}_\alpha$; $Z^\alpha\, \partial/\partial Z^\alpha$; $\mathrm{i}(\partial/\partial Z^\alpha) \otimes (\partial/\partial \bar{Z}_\alpha) - \mathrm{i}(\partial/\partial \bar{Z}_\alpha) \otimes (\partial/\partial Z^\alpha)$; $\mathrm{i}\mathrm{d}Z^\alpha \wedge \mathrm{d}\bar{Z}_\alpha = \omega$ are all preserved in the sense that $\delta(Z^\alpha \bar{Z}_\alpha) = 0$; $\delta(Z^\alpha \mathrm{d}\bar{Z}_\alpha) = 0$; $\delta \circ Z^\alpha\, \partial/\partial Z^\alpha = Z^\alpha\, \partial/\partial Z^\alpha \circ \delta$ and so on. If we define

$$[\chi, \Psi] := -\mathrm{i}\,\frac{\partial \chi}{\partial Z^\alpha}\,\frac{\partial \Psi}{\partial \bar{Z}_\alpha} - \mathrm{i}\,\frac{\partial \chi}{\partial \bar{Z}_\alpha}\,\frac{\partial \Psi}{\partial Z^\alpha} \tag{3.25}$$

then

$$\delta \Psi = [\Psi, H] \tag{3.26}$$

$$-\delta(\mathrm{d}Z^\alpha) = \mathrm{d}\left(\mathrm{i}\,\frac{\partial H}{\partial \bar{Z}_\alpha}\right) = \mathrm{i}\,\frac{\partial^2 H}{\partial Z^\beta \partial \bar{Z}_\alpha}\,\mathrm{d}Z^\beta + \mathrm{i}\,\frac{\partial^2 H}{\partial \bar{Z}_\beta \partial \bar{Z}_\alpha}\,\mathrm{d}\bar{Z}_\beta \tag{3.27}$$

and from these one can check the invariances mentioned above.

If we consider any weak gravitational wave of any shape whatever, which separates two regions of flat space-time, then we are led to equations of exactly similar form to the above. This is because weak gravitational waves can be superposed linearly and can be broken down into a superposition of waves of the above types only. (Actually plane waves alone will suffice for this.) The corresponding H functions are likewise linearly composed of those above.

We must define what we mean by ϕ, ω on the N associated with a general curved space. In flat space we have seen that (2.22, 2.23)

$$\phi \equiv \mathrm{i}Z^\alpha \mathrm{d}\bar{Z}_\alpha = P_a \mathrm{d}x^a \tag{3.28a}$$

$$\omega \equiv \mathrm{i}\mathrm{d}Z^\alpha \wedge \mathrm{d}\bar{Z}_\alpha = \mathrm{d}P_a \wedge \mathrm{d}x^a = \nabla_{[a}P_{b]}\,\mathrm{d}x^a \wedge \mathrm{d}x^b \tag{3.28b}$$

where for the right hand expressions of (3.28b) we take P_a to be the tangent vector field of a congruence of geodesics. In curved space we use these as definitions of ϕ, ω. This is possible because $P_a \mathrm{d}x^a$ and $\mathrm{d}P_a \wedge \mathrm{d}x^a$, as forms applied to connecting vectors of neighbouring geodesics, are constant (the constancy in this sense of $\mathrm{d}P_a \wedge \mathrm{d}x^a$ being the well-known Lagrange identity, see [26, 27]) and since N (modulo the phase factors) has been identified with the space of null geodesics the forms ϕ, ω will be invariantly defined on N.[22] The expressions (3.28) lead to the interpretations of ϕ and ω mentioned before as respectively, time-displacement and rotation of neighbouring null geodesics.

---

[22] See footnote on the next page.

In both the examples A and B above we find, provided Einstein's vacuum equations hold, that we may write

$$H(Z^\alpha, \bar{Z}_\alpha) = H^+ + H^-; \qquad H^+ = \overline{H^-}$$

$$H^+ = \bar{Z}_\alpha I^{\alpha\beta}\, \partial g/\partial Z^\beta \tag{3.29}$$

where $g(Z^\alpha)$ is holomorphic and homogeneous of degree 2 in Z^α. Explicitly, for the case A, we have

$$H^+ = |Z^2|^2 \bar{r}(-\mathrm{i}Z^1/Z^2) = \bar{Z}_\alpha I^{\alpha\beta} \partial(g(Z^\alpha))/\partial Z^\beta, \tag{3.30}$$

so we obtain (3.29) if

$$g(Z^\alpha) = \mathrm{i}(Z^2)^2 \int_{x_0}^{-\mathrm{i}Z^1/Z^2} \bar{r}(x)\mathrm{d}x.$$

The infinity twistor appears in (3.29) because gravitation is not conformally invariant: $I^{\alpha\beta}$ is the conformal-symmetry breaking term which tells us "where" infinity is.

We can similarly treat electromagnetic scattering, introducing charged zero-rest-mass particles, with momentum P^a. The acceleration of such a particle is

$$P^a \nabla_a P^b = eF^{ab} P_a \tag{3.31}$$

where e is the charge and F^{ab} the Maxwell field. This gives well-defined equations of motion for the particle even though its rest-mass is zero. We can consider an idealised situation similar to that of the gravitational impulse waves considered above. Here we take two regions of field-free space separated by an electromagnetic plane or spherical wave of δ-function amplitude. A zero-rest-mass particle on either side of the wave may be described by a null twistor. The wave imparts an impulse to the particle and so defines a transformation of the twistor space from one side to the other. The transformation again turns out to be of Hamiltonian form, but now, in the infinitesimal case H turns out to be homogeneous of degree zero separately in Z^α and in $\bar{Z}_\alpha$, where $H = H^+ + \bar{H}^+$ with $H^+ = f(Z^\beta)$ holomorphic and of degree zero in Z^α (assuming F^{ab} satisfies the free-space Maxwell equations). The treatment may likewise be extended to any infinitesimal scattering by linear superposition of such waves, and hence of the corresponding H functions.

We have encountered holomorphic functions in both the gravitational and electromagnetic cases.

---

[32] Footnote from preceding page. The invariance of ω is closely related to some general theory concerning symplectic manifolds (i.e. manifolds with a two-form W of maximal rank such that $\mathrm{d}W = 0$). Let $\mathcal{T}$ be any manifold, with a symplectic structure W, e.g. the cotangent bundle $\mathcal{T} = T^*(\mathcal{M})$ of a manifold $\mathcal{M}$, with $W = \mathrm{d}P_a \wedge \mathrm{d}x^a$, where P_a denotes a cotangent vector at x [26]. Then on any hypersurface S in $\mathcal{T}$, W must be degenerate (S being odd-dimensional), and so at each point of S there is a vector which W maps to zero. If we factor out by equivalence along the integral curves of this vector field, the new manifold (of equivalence classes) again has a symplectic structure (with two less dimensions than that of $\mathcal{T}$). This is assuming that the "normal" situation obtains, whereby the factoring procedure actually produces a smooth manifold (of two less dimensions than $\mathcal{T}$).

In the case of twistor space we may take $T^*(V_4)$, the cotangent bundle of space-time, and consider the seven-dimensional subspace S given by $P^a P_a = m^2 = 0$. The equivalence classes are the null geodesics, and the resulting six-dimensional space G can also be obtained by applying the same procedure to $\mathcal{T} = \mathrm{C}$, the hypersurface S now being N and the vector field on N generating $Z^\alpha \to e^{\mathrm{i}\theta} Z^\alpha$. These transformations leave the null geodesic Z and its tangent vector $\bar{\pi}^A \pi^{A'} = P^a$ unaffected, so in either case we obtain G as the space of affinely parametrised null geodesics; its symplectic structure being induced by $\mathrm{d}P_a \wedge \mathrm{d}x^a$ in the first case, or equivalently, by $\mathrm{i}\mathrm{d}Z^\alpha \wedge \mathrm{d}\bar{Z}_\alpha$ in the second.

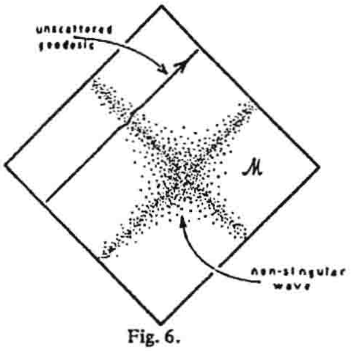


Fig. 6.

But in general these have poles, and if not, then H is a bilinear function of Z^α, $\bar{Z}_\alpha$. For the fields of this latter simple type the null geodesics emerge ultimately unscattered although waves can come in and go out (fig. 6). However, in general cases where the wave possesses singularities, singularities in the function can exist and scattering occurs. For instance in example B above one can see from the fact that there is no non-singular non-constant harmonic function on a 2-sphere that the behaviour of $f(\zeta)$ on the null cone joining the two flat spaces must be singular on at least one generator of the cone. This singularity could be cancelled by putting together an appropriate set of such cones but we would then be back to the situation of fig. 6 where there is no scattering.

3.3. *Quantization*

We wish to pass from the scattering of zero-rest-mass particles by a (weak) gravitational or electromagnetic wave to the scattering of zero rest mass fields. In general a zero rest mass field $\phi_{A'B'\ldots L'}$ is defined by a holomorphic function $f(Z^\alpha)$. We may now ask how to transform f in order to represent the scattering of the $\phi_{\ldots}$ field. A somewhat formal answer is provided if we use the correspondence $\bar{Z}_\alpha \to -\partial/\partial Z^\alpha$ suggested by the fact that Z^α and $\bar{Z}_\alpha$ are canonically conjugate variables. Thus, we write

$$H(Z^\alpha, -\partial/\partial\bar{Z}^\alpha) \text{ for } H(Z^\alpha, \bar{Z}_\alpha),$$

and apply it to f. We are here regarding H as describing the effect of a fixed given gravitational field. Now with the H considered above for the scattering of massless particles by weak gravity we get

$$H^+ \to -\frac{\partial}{\partial Z^\alpha} \circ I^{\alpha\beta} \frac{\partial g}{\partial Z^\beta} = -\left(I^{\alpha\beta} \frac{\partial g}{\partial Z^\beta}\right) \frac{\partial}{\partial Z^\alpha} \tag{3.32}$$

the commutation being possible because $I^{\alpha\beta}$ is skew. Thus no factor ordering problem arises. Similarly we would have

$$H^- \to Z^\alpha I_{\alpha\beta} [\partial\bar{g}/\partial Z_\beta]_{\bar{Z}_\beta \to -\partial/\partial Z^\beta}$$

which is more awkward! However we aim to consider matrix elements $\langle g|H|f\rangle$ and therefore need not evaluate $H^-|f\rangle$ as such, for

$$\langle g|H|f\rangle = \langle g|H^-|f\rangle + \langle g|H^+|f\rangle$$

and we may take H^- to act on $\langle g|$ writing $H^- \to (\overleftarrow{\partial}/\partial\bar{Z}_\alpha) I_{\alpha\beta}\, \partial\bar{g}/\partial\bar{Z}_\beta$. So far we have not defined what we mean by $\langle g|f\rangle$ and our next task is therefore to set up a Hilbert space of functions f. In doing so we can be guided by the need for suitably nice formal formulae and agreement with the scalar product used by Fierz [30].

From $\phi_{A\ldots L}$ we may construct a series of potentials $\overset{(k)}{\phi}{}^{\overbrace{A'B'\ldots D'}^{k\text{ indices}}}_{EF\ldots L}$ satisfying

$$\nabla_{AA'}\overset{(1)}{\phi}{}^{A'}_{B\ldots L} = \phi_{A\ldots L}$$

$$(3.33)\quad \left\{\begin{aligned} &\nabla_{DD'}\overset{(k)}{\phi}{}^{A'\ldots D'}_{E\ldots L} = \overset{(k-1)}{\phi}{}^{A'\ldots C'}_{D\ldots L} \\ &\nabla^{EE'}\overset{(k)}{\phi}{}^{A'B'\ldots D'}_{EF\ldots L} = 0 \end{aligned}\right.$$

$$\nabla^m\nabla_m\overset{(n)}{\phi}{}^{A'B'\ldots L'} = 0.$$

At each step there is a gauge freedom in choice of $\overset{(k)}{\phi}{}_{\ldots}$ cf. [4]. Following Fierz, we may now define (with suitable numerical constant k)

$$\langle\chi|\phi\rangle := k\int_S \overset{(n-1)}{\phi}{}^{A'B'\ldots K'}_{L}\,\chi_{A'B'\ldots L'}\,\mathrm{d}S^{LL'} \tag{3.34}$$

where S is a spacelike surface. One must (and can) check that this is gauge independent, independent of the choice of surface, that one may interchange χ and ϕ yielding a Hermitian symmetry, and that the product is conformally invariant (cf. [36]).

Our next task is to express the scalar product in terms of $f(Z^\alpha)$, $g(W_\alpha)$. Physically meaningful answers must be contour integrals since if f is replaced by f' where $f - f'$ is nonsingular inside the integration contour of eq. (2.20) the field is not changed. Let us investigate the form that such a contour integral must take. Suppose we have $\beta(Z^\alpha, X^\alpha, \ldots, W_\alpha \ldots)$ which is a function homogeneous of degree (-4) in each variable. To integrate one must define a differential form $\mathcal{D}ZX..W\ldots$, so that the integral depends only on the homology class (relative to the space less regions of singularity) of the region of integration (i.e. we require that the resulting object is a genuine contour integral). For this we use

$$\left.\begin{aligned} \mathcal{D}Z &= \tfrac{1}{6}\epsilon_{\alpha\beta\gamma\delta}Z^\alpha \mathrm{d}Z^\beta \wedge \mathrm{d}Z^\gamma \wedge \mathrm{d}Z^\delta \\ \mathcal{D}W &= \tfrac{1}{6}\epsilon^{\alpha\beta\gamma\delta}W_\alpha \mathrm{d}W_\beta \wedge \mathrm{d}W_\gamma \wedge \mathrm{d}W_\delta \end{aligned}\right\} \tag{3.35}$$

$$\mathcal{D}ZX..W\ldots = \mathcal{D}Z \wedge \mathcal{D}X \wedge ..\mathcal{D}W \wedge \ldots.$$

We then find that $\mathrm{d}(\beta\mathcal{D}ZX..W..) = 0$ as required for $\oint \beta\mathcal{D}ZX..W..$ to be dependent on the homology class of the contour, and that the integral is a scalar, as we desire [21].

To illustrate the value of such contour integrals let us digress for a moment. If we take $\beta(Z^\alpha)$ to represent an electromagnetic field, then $\oint \beta(Z^\alpha)\mathcal{D}Z$ gives the charge integral for a source for the field. In a gravitational field the function $f(Z^\alpha)$ introduced before is of degree -6, so we take

$$\oint Z^{\alpha}Z^{\beta}f(Z)\,\mathcal{D}Z$$

and find this is the twistor describing the energy momentum and angular momentum of a source for the field.

The same differential forms are now used for the scalar products in terms of f and g. We must insert additional factors so that $f(Z^{\alpha})g(W_{\alpha})$ has the correct homogeneity degree i.e. $(-4, -4)$. For $n = 0, 1$ this may be done by[23]

$$\oint \underbrace{f(Z^{\alpha})}_{\text{degree}-n-2}\quad \underbrace{g(W_{\alpha})}_{\text{degree}-n-2}\quad (W_{\beta}Z^{\beta})^{n-2}\,\mathcal{D}ZW \tag{3.36}$$

since there is then a 6-dimensional contour not homologous to zero in the 16-dimensional subspace of $C \otimes C^{*}$ where $f, g, (Z^{\alpha}W_{\alpha})^{n-2}$ are non-singular (i.e. there is a contour surrounding the singularities). However for $n \geqslant 2$ this formula is no longer satisfactory since the $(Z^{\alpha}W_{\alpha})^{n-2}$ singularity disappears. If we consider defining successive factors (as n increases) by integrating the previous ones, the factor for $n = 0$ will be $\log(W_{\alpha}Z^{\alpha})$, which is not homogeneous. So we take

$$(x)_k = (-x)^{-k}\Gamma(k) \tag{3.37}$$

for non-integer k and formally define the values at integers by taking the obvious limit. The point is that for $k = -1, -2, \ldots$ the singular behaviour of $\Gamma(k)$ compensates for the lack of a pole $(-x)^{-k}$. Thus in (3.36) for general n we write $(W_{\alpha}Z^{\alpha})_{2-n}$ as defined by (3.37) in place of $(W_{\alpha}Z^{\alpha})^{n-2}$. (In fact for $n = 0$ we are led to

$$\log\{(W_{\alpha}Z^{\alpha})/(W_{[\alpha}C_{\beta}D_{\gamma]}Z^{\alpha}A^{\beta}B^{\gamma})\} \tag{3.38}$$

and one does then find an answer which is independent of the auxiliary twistors $A^{\gamma} \ldots D_{\alpha}$.) With these definitions it can be checked that the basis functions used in (2.21) are orthogonal. The functions $(W\cdot Z)_k$ do satisfy the formal property

$$\partial(W\cdot Z)_k/\partial Z^{\alpha} = W_{\alpha}(W\cdot Z)_{k+1} \tag{3.39}$$

which in fact is what is really used in actual calculations.

4. Evaluation of scattering amplitudes

4.1. *Introduction*

In this section we consider the evaluation of scattering amplitudes for basic processes involving zero rest-mass particles. It turns out that the twistor formalism suggests that certain types of contour integral be performed, each contour integral being associated with a graph of a particular type. In the case of electromagnetic scatterings the twistor calculations give agreement with the standard results for simple processes (e.g. Möller, Compton scattering) but a general theory is still lacking which relates the twistor approach explicitly with the normal Feynman rules. Also, the correct treatment of rest-mass and of gravitational interactions within the twistor formalism has not yet emerged. Of necessity the treatment given here will be somewhat sketchy and to some extent unmotivated. A more complete treatment will have to await a more complete theory.

---

[23] The scalar product of two functions $f(Z^{\alpha})$ and $g(\bar{W}^{\alpha})$ is defined to be zero unless they have the same spin and the same sign of frequency.

The basic idea here will be to use the twistor description of massless fields in terms of holomorphic functions (cf. section 2.3) in order to represent the in- and out-states of a scattering process. These in- and out-states should, strictly speaking, be wave packets, in this theory, rather than pure momentum states. This leads to certain difficulties in making comparisons with existing theory, where pure momentum states are almost invariably used. However, a formal method for treating momentum states within the twistor theory will also be given.

The wave packets defining the in- and out-states can be described by (normalized) solutions of the zero mass spin s equations: $\nabla^{AP'}\phi_{AB...L} = 0$, $\nabla^{PA'}\theta_{A'B'...L'} = 0$. Since these fields are obtained from the twistor holomorphic functions $f(Z)$ by means of contour integrations and they do not define the f's uniquely, we must expect that any scattering amplitude which is a functional of the f's, must itself be obtainable from the f's by means of some form of contour integration. Now we have seen an example – albeit a trivial one – of a scattering amplitude already, namely the scalar product $\langle\alpha|\beta\rangle$. This expresses the amplitude that the out-state is α... given that the in-state is β..., where the particle enters and leaves without interacting. The corresponding twistor contour integral is

$$\oint a(W)b(Z)\,(W\cdot Z)_{2-n}\,\mathcal{D}WZ \tag{4.1}$$

(cf. (3.37)).

In order to motivate, to some extent, the generalizations of this to include interactions, it will be worthwhile briefly to examine the scatterings we considered in the last section by a classical gravitational or electromagnetic wave. Recall that the Hamiltonian $H = H^+ + H^-$ for a gravitational scattering was defined by

$$H^+ = \overline{H^-} = I^{\alpha\beta}\bar{Z}_\alpha\,\partial g/\partial Z^\beta \tag{4.2}$$

where g is a holomorphic function of Z^α which is homogeneous of degree +2, whereas for an electromagnetic scattering we had

$$H^+ = f \tag{4.3}$$

where f is holomorphic in Z^α and homogeneous of degree 0. These homogeneity degrees may be compared with –6 and –4, respectively, which are the homogeneity degrees of those holomorphic functions which, upon contour integration, yield a linearized gravitational or electromagnetic field (cf. section 2.3). We may think of these latter holomorphic functions (the negative degree functions) as describing the fields in a "passive" capacity, whereas the holomorphic functions (non-negative degree) which appear in the Hamiltonian are "active" in that they effect scattering of other fields. We must expect that there should be some way of passing from a "passive" function to an "active" one. A formula which achieves this must have the effect of increasing the homogeneity degree by four in the electromagnetic case and by eight in the gravitational case.

Now consider the expressions (with numerical factors chosen for convenience)

$$g(Z^\alpha) = \frac{1}{(2\pi i)^3}\oint p(W_\alpha)(W_\beta Z^\beta)_{-2}\,\mathcal{D}W \tag{4.4a}$$

$$f(Z^\alpha) = \frac{1}{(2\pi i)^3} \oint q(W_\alpha)(W_\beta Z^\beta)_0 \mathcal{D}W \tag{4.4b}$$

where p and q have respective degrees of homogeneity -6 and -4 as is required for "passive" gravitational and electromagnetic functions. As a consequence the functions g and f have the required "active" degrees $+2$ and 0, respectively. To some extent the expressions (4.4) must be regarded as merely formal, however. The reason for this is concerned with the nature of the singularity sets of p and q. Recall that in order to define a zero rest-mass field, we required not merely a holomorphic function, but also a division of its singularity sets in C^+ (or in C^-) into two disconnected regions, the contours being chosen so as to separate one region from the other. But this is not adequate for defining the contours for (4.4). Instead, one needs, for topological reasons, a division of the singularity sets of p and q into *three* disconnected regions[24]. Thus more information is required, in order to evaluate g and f, than just that which is needed to evaluate the zero rest-mass fields from the functions p and q. Having made this observation, however, we may regard eqs. (4.4) as being valid in a certain sense. This may be verified explicitly by reference to particular functions p, q of the form $p(W_\alpha), q(W_\alpha) = 1/(W_\beta E^\beta)(W_\gamma F^\gamma)(W_\delta G^\delta)^n$ (with $G^\alpha I_{\alpha\beta} = 0$; $n = 4,2$) which represent certain plane waves. The resulting $g(Z^\alpha), f(Z^\alpha) \propto (E^\alpha F^\beta K^\gamma Z^\delta \epsilon_{\alpha\beta\gamma\delta})^{n-1}/E^\lambda F^\mu G^\nu Z^\tau \epsilon_{\lambda\mu\nu\tau}$ (K^γ arbitrary) are essentially the correct "active" functions for this case. The general case may then be viewed as arising from linear combinations of such plane waves. But difficulties remain if we desire to make (4.4) rigorous.

Let us now attempt to use (4.4) in order to express the scattering of a zero rest-mass field by a gravitational field defined by p or by an electromagnetic field defined by q. According to section 3.3 we may expect to describe this scattering by means of a matrix $\langle\alpha|H^+|\beta\rangle$ or $\langle\alpha|H^-|\beta\rangle$. Substituting (4.2) or (4.3) (with $\bar{Z}_\alpha \to -\partial/\partial Z^\alpha$ or $Z^\alpha \to \partial/\partial\bar{Z}_\alpha$) into (4.1) and using (4.4) we arrive at expressions of the form (again choosing numerical factors for convenience)

$$\frac{1}{(2\pi i)^5} \oint a(W)(W_\gamma Z^\gamma)_{2-n} I^{\alpha\beta} \frac{\partial}{\partial Z^\alpha} b(Z) \frac{\partial}{\partial Z^\beta} \{(U_\delta Z^\delta)_{-2}\, p(U)\}\, \mathcal{D}WZU$$

$$= -\frac{1}{(2\pi i)^5} \oint a(W)\, b(Z)\, p(U)\, (W{\cdot}Z)_{3-n}\, (U{\cdot}Z)_{-1} (I^{\alpha\beta} W_\alpha U_\beta) \mathcal{D}WZU$$

and

$$\frac{1}{(2\pi i)^5} \oint a(W)\, b(Z)\, q(U)\, (W{\cdot}Z)_{2-n} (U{\cdot}Z)_0\, \mathcal{D}WZU$$

respectively.

It will be useful to represent the singularity structure of the integrands of such contour integrals in a graphical way. The twistor variables which are to be integrated over will be represented by vertices, a black vertex denoting an upper index twistor (e.g. Z^α) and a white vertex a lower index twistor (e.g. W_α). A line joining two vertices denotes the singularity corresponding to the vanishing of the scalar product of the two twistors representing the two vertices. Thus lines are allowed joining vertices only if one is black and the other white. In addition, lines will be allowed which have a

---

[24] In fact the functions considered in section 5.2 do not suffer from this fault.

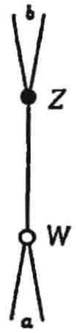


Fig. 7.

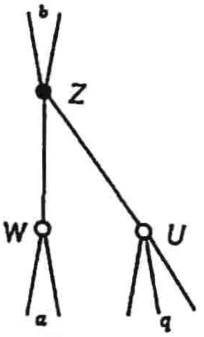


Fig. 8.

vertex at one end only. Such lines represent the singularities in the holomorphic functions $a(Z)$, $b(Z)$, ...,... which enter into the integrals. In general the singularities of b will be separated into two sets, so two corresponding lines would emerge from the vertex representing Z, each of which stands for one set of singularities of b. We may, in fact, think in terms of particular functions b of the form

$$b = (A{\cdot}Z)^{-1-r}(B{\cdot}Z)^{-1-n+r}. \tag{4.5}$$

Then we may regard each of the lines with just one vertex as representing one of the two regions $A{\cdot}Z = 0$, $B{\cdot}Z = 0$. Thus the constant twistors A_α, B_α are represented by the missing vertices at the ends of the two lines. We call states describable by such functions (4.5) *elementary states.*

In fig. 7 the diagram representing the singularity structure for the scalar product (4.1) is depicted; the diagram in fig. 8 is what we have been led to for the singularity structure arising for our proposed interaction integrals. We can make our diagrams more explicit by labelling the lines with integers, where a single line labelled by r stands for a factor $(W{\cdot}Z)_{r+1}$. This can apply whether or not vertices appear at each end of the line, so that elementary states can be explicitly incorporated into the diagrams. For example, the scalar product between the two elementary states given

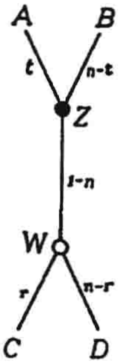


Fig. 9.

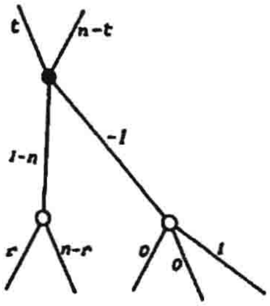


Fig. 10.

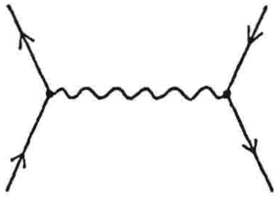


Fig. 11.

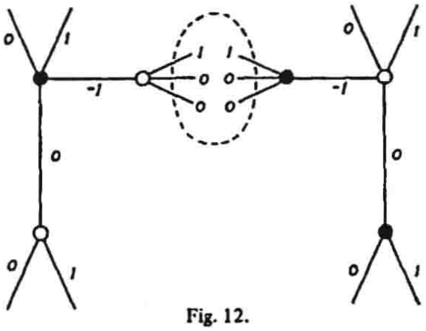


Fig. 12.

by $a(W) = (W\cdot C)_{r+1}(W\cdot D)_{n-r+1}$, $b(Z) = (A\cdot Z)_{t+1}(B\cdot Z)_{n-t+1}$, namely

$$\frac{1}{(2\pi i)^5}{}'\oint (W\cdot C)_{r+1}(W\cdot D)_{n-r+1}(W\cdot Z)_{2-n}(A\cdot Z)_{t+1}(B\cdot Z)_{n-t+1}\mathcal{D}WZ$$

can be explicitly expressed by the diagram of fig. 9. Similarly (restricting attention to the electromagnetic case only) the proposed interaction integral could be depicted by the diagram of fig. 10, the electromagnetic field being described by the function $q(Z) = (E\cdot Z)_1(F\cdot Z)_1(G\cdot Z)_2$.

We may think of fig. 10 as describing some sort of electromagnetic vertex analogous to the one arising in standard Feynman theory. Then we could attempt to build up the diagrams corresponding to simple processes such as Möller or Compton scattering, for example.

Consider Möller scattering, the standard Feynman diagram being that of fig. 11 (except that here the electrons must be massless). We have been led to consider a twistor diagram of the sort indicated in fig. 12 where the arms in the dotted circle are to be "summed over" or contracted together in some other appropriate way. Because of the fact that the formula depicted diagrammatically in fig. 13 can be shown to be valid, it appears reasonable to perform our contraction so as to obtain fig. 14.

Now recall that we have employed the decomposition $H = H^+ + H^-$ in obtaining this diagram. Were we to repeat the discussion with H^- in place of H^+, we should end up with the diagram of fig. 15. It may seem to be plausible to associate the twistor diagrams of figs. 14 and 15 with processes depicted in figs. 16a and 16b respectively, where the arrows denote helicity (in units of $\frac{1}{2}\hbar$) so the double arrow on the photon line occurs because the spin is 2 × $\frac{1}{2}\hbar$. We might then imagine that (corresponding to $H = H^+ + H^-$) the Möller scattering, fig. 11, could be expressible as the sum of two terms: fig. "16a plus 16b" or fig. "14 plus 15". However this presupposes that it is meaningful to express a virtual photon as a sum of two parts each of which has a well-defined helicity. This possibility is denied, in the conventional theory, owing to the fact that a virtual photon must be

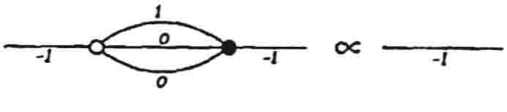


Fig. 13.

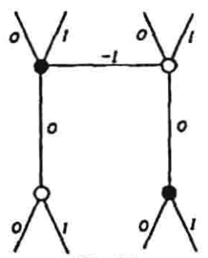


Fig. 14.

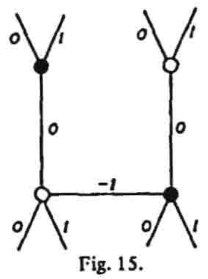


Fig. 15.

allowed to be "off the mass shell", i.e. to have a non-zero (and sometimes imaginary) rest-mass. Particles with non-zero rest-mass do not have a well-defined helicity. In the twistor approach, on the other hand, it would be unreasonable to allow photons to have a rest-mass, since this would go against the basic philosophy of the theory. Nevertheless the theory could not give sensible answers (for Möller scattering, for example) if it did not in some way reflect the fact which, in the conventional formalism, is accounted for by allowing virtual photons to be off the mass shell. This, in itself, renders it unlikely that the twistor computation of Möller scattering could be obtainable as a sum of two integrals, like those represented in figs. 14 and 15, in each of which the contribution due to a virtual photon appears to be identifiable, the photon having a well-defined helicity[25].

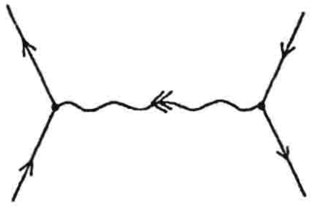


Fig. 16a.

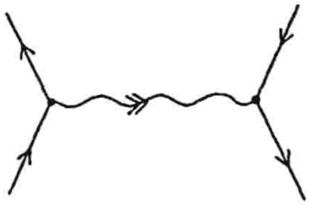


Fig. 16b.

In fact this is borne out by the actual computation of the integrals for figs. 14 and 15. It turns out that the answer in each case is simply proportional to that which it would have been, had the segment bearing "−1" been omitted. That is to say there is no interaction expressed by these integrals. On the other hand, we may envisage "superimposing" the two figs. 14 and 15. The result is fig. 17 and, remarkably, the integral it expresses actually gives, as we shall see later, the correct angular dependence for Möller scattering for massless electrons (i.e. the high energy limit for massive electrons) when the helicities of the two incoming particles are opposite (as indicated in fig. 11). It should be emphasized that at this stage there is no theoretical basis for superimposing figs. 14 and 15. On the other hand, no other reasonable way of modifying fig. 14 to obtain any scattering at all – let alone the correct Möller scattering – presented itself. The essential correctness of fig. 17 in representing a single photon exchange is further substantiated by the fact that it leads

---

[25] We are grateful to B.S. DeWitt for drawing our attention to this fact.

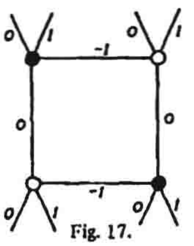


Fig. 17.

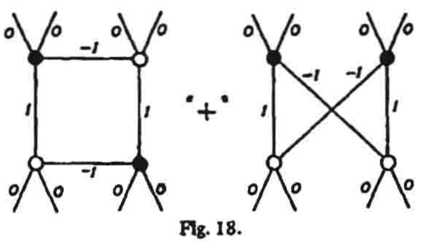


Fig. 18.

naturally to a similar expression (fig. 18) which correctly gives the angular dependence for the case of two spin zero massless particles exchanging a photon. Here we must add the results for two diagrams: in the first the spinless particles have opposite "helicities" and in the second, the same "helicity". (It appears that, in the present theory, two separate "helicities" must be distinguished for a massless particle even when the spin is zero, but to obtain the physical answer, these "helicities" must be summed over.) It may be remarked at this point that recent work based on the results of section 5.3 has subsequently fully justified these twistor expressions. Finally, we shall also see later that fig. 17 leads directly to another diagram (fig. 30) which correctly gives the angular dependence for the high energy limit (i.e. massless particles) for Compton scattering.

4.2. *Rules for diagrams*

Accepting, for the moment, the correctness of fig. 17, we are led[26] to suggest some additional rules for the construction of our diagrams. First, let us insist that exactly *four* lines enter each vertex. In order to depict the scalar product (cf. fig. 7), we shall then require a *double* line for the internal segment. We adopt the convention that a single integer is to be attached to the double line as a whole. The line then stands for the factor $(W_\alpha Z^\alpha)_{r+2}$, where the twistors W_α and Z^α correspond to the vertices at the two ends of the line and r the integer attached to it. Similarly we shall occasionally need to use a triple line in a diagram, standing for the factor $(W_\alpha Z^\alpha)_{r+3}$. For completeness, a quadruple line standing for $(W_\alpha Z^\alpha)_{r+4}$ will also be allowed, but because of the rule that just four lines can terminate at each vertex, such lines must be disconnected from any other portions of the diagram. (Thus they would appear, if anything, to stand for disconnected vacuum processes.) In fact we must have $r = 0$ in this case because, in order for a diagram, constructed according to the above rules, to represent an integral which is homogeneous of degree zero (necessary if the integral is to be a genuine contour integral, invariant under continuous deformations of the contour), then we must require that: *the sum of the integers on all lines entering a given vertex is zero.* The notation for the diagrams is summarized in fig. 19.

The diagram for the scalar product can now be drawn as in fig. 20. Here the in- and out-states are explicitly expressed as the elementary states defined by $(W_\alpha A^\alpha)_1 (W_\beta B^\beta)_{n+1}$ and $(C_\alpha Z^\alpha)_1 (D_\beta Z^\beta)_{n+1}$ respectively. These elementary states have particular interest in that the fields are everywhere null, see section 2.3. If $n = 4$ and A^α, C_α are null twistors then they are linearized Robinson–Trautman null waves [31] of a particular kind, having wavefronts which are null cones with vertices on the null line A. If A and B are both at infinity we have a constant plane wave, while if A alone is at in-

---

[26] There are also some topological reasons, concerning the contour surfaces, for preferring diagrams constructed according to these rules. They are not conclusive, however.

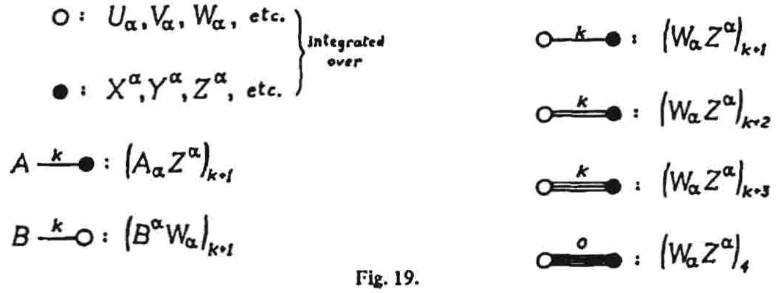

○ : $U_\alpha, V_\alpha, W_\alpha$, etc.
● : $X^\alpha, Y^\alpha, Z^\alpha$, etc.
integrated over
A —k— ● : $(A_\alpha Z^\alpha)_{k+1}$
B —k— ○ : $(B^\alpha W_\alpha)_{k+1}$
○—k—● : $(W_\alpha Z^\alpha)_{k+1}$
○═k═● : $(W_\alpha Z^\alpha)_{k+2}$
○≡k≡● : $(W_\alpha Z^\alpha)_{k+3}$
○—0—● : $(W_\alpha Z^\alpha)_4$

Fig. 19.

finity we have a plane wave of non-constant amplitude. In all these particular cases the fields possess singularities (possibly only at infinity); but if we require that in the C-picture the line AB or C ∩ D lie entirely in C⁺ [resp. C⁻] then we shall have a nowhere singular field which has positive [resp. negative] frequency (cf. section 2.3), cf. [4].

However, we shall not always want to represent our in- and out-states as elementary states. For more general situations (e.g. plane sinusoidal waves) it ceases to be useful to label the singularity sets of $a(z)$, ... by two different integers; only their sum (= twice the spin) has significance. For this reason, we shall frequently represent an in- or out-state by one *double* external line, to which we attach merely a single integer n (this being twice the spin of the field). In fig. 21 we have drawn the scalar product according to these conventions. We should keep in mind, however, that the status of the internal double line (marked $-n$) is really rather different from that of the external double lines. The meaning of internal double lines has to do with the topological set-up with regard to the contours. Roughly speaking, we have to surround each singularity denoted by a single line by an S^1 (topological circle), each singularity denoted by an internal double line by an S^3 (topological 3-sphere) and each singularity denoted by an internal triple line by an S^5, taking the appropriate (twisted?) topological product. Then, in effect, to get the dimension up or down to the required value for the contour (i.e. to $3v$ real dimensions, where v is the number of vertices) we must "multiply" or "divide" by the appropriate number of S^1's. (This means that for the quadruple line we integrate over a $\mathbb{CP}^3 \cong S^7/S^1$.)

It appears to be consistent to regard a double line as a pair of single lines collapsed together in a certain sense. The equation represented by fig. 22 is in fact valid provided we interpret the two separated lines which join the same two vertices (on the left-hand side of the equation) in an appropriate way. We cannot simply adopt the product of factors $(W_\alpha Z^\alpha)_{1-a}(W_\beta Z^\beta)_{b+1}$ in the integral, since the contour is supposed to pass "between" the singularity represented by each factor. Since

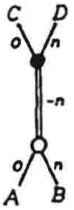


Fig. 20.

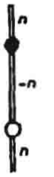


Fig. 21

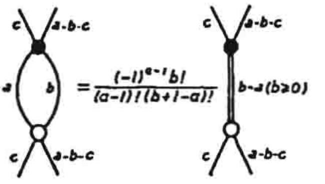

c
a-b-c
-a
b
$= \frac{(-1)^{a-1}b!}{(a-1)!(b+1-a)!}$
c
a-b-c
b-a (b≥0)
c
a-b-c
c
a-b-c

Fig. 22.

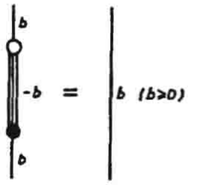


Fig. 23.

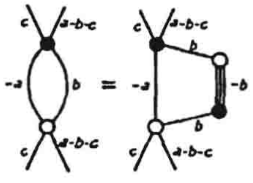

c
a-b-c
-a
b
=
c
a-b-c
b
-a
-b
b
c
a-b-c
c
a-b-c

Fig. 24.

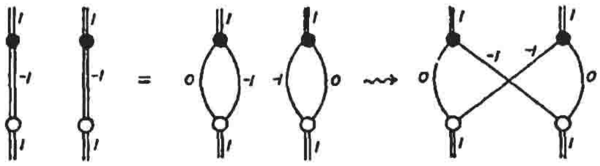


Fig. 25.

the two factors here have the *same* singularity (namely $W_\alpha Z^\alpha = 0$) it is clearly impossible for the contour to pass between them. However we can invoke the relation expressed in fig. 23 in order to "free" these singularities from one another. The meaning of the left-hand side of fig. 22 is then expressed in fig. 24, for which a perfectly well-defined contour does exist. We should finally remark that in order to get the numerical factors simple in these expressions[27] it is convenient to adopt a suitable multiplying factor before the integral representing each diagram. We shall take this factor to be $(2\pi i)^{-3v+r}$ where v is the number of vertices and r is the number of extra lines which have to be drawn which make the multiple lines multiple (i.e. the smallest number of lines which must be erased so that no multiple lines remain). Thus the integral for the right-hand side of fig. 22 will be preceded by the factor $(2\pi i)^{-5}$, that for the left-hand side of fig. 23 by $(2\pi i)^{-4}$, and that for the right-hand side of fig. 24 by $(2\pi i)^{-10}$.

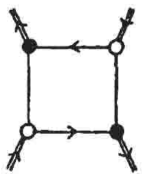


Fig. 26.

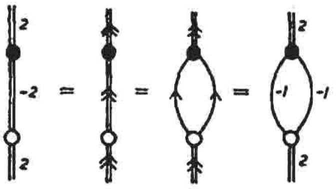


Fig. 27.

It is worth remarking here on a curious way of interpreting the electromagnetic interaction expressed in fig. 17. Let us imagine two charged zero-mass spin $\frac{1}{2}$ particles of positive helicity coming together. If each particle emerges without having interacted, then the twistor diagram is that depicted at the far left in fig. 25, namely two scalar product diagrams multiplied together. According to fig. 22, we can re-express the situation by the diagram in the centre of fig. 25. Still there is no interaction. But if we imagine that the two lines marked "-1" are subject to some form of Fermi (or Bose) statistics, then we must expect to have to add or subtract a contribution in which the lines marked "-1" cross over – as depicted at the extreme right of fig. 25. But this resulting figure is just that which we have used, in fig. 17, to denote a single photon exchange between the two particles (except that the photon lines are now depicted as crossing over – simply because the helicities of the two incoming particles have been chosen to be the same, rather than opposite). Thus, the twistor formalism seems to be suggesting that the electromagnetic interaction may be some form of manifestation of Fermi (?) statistics as applied to the individual twistor scalar product lines which occur in the diagrams.

---

[27] We are grateful to A. Qadir for evaluating several of these integrals in detail [32].

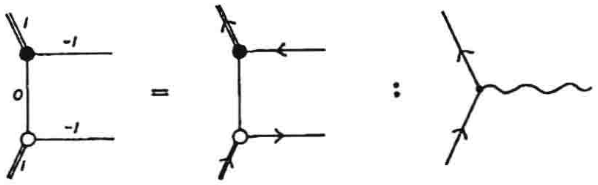


Fig. 28.

The integers which occur on the lines in the twistor diagrams may be interpreted as denoting helicity. Thus the integer r appearing on a line indicates that a helicity $\frac{1}{2}r$ is proceeding from the black vertex to the white vertex which terminate the line. The fact that the integers on all the lines terminating at any one vertex must sum to zero gives us a *helicity conservation law* for diagrams constructed in this way: the total helicity entering the diagram must equal the total helicity leaving it. In fig. 26 the helicities involved in the photon exchange of fig. 17 are indicated. Note that the exchanged photon is represented by two disconnected lines, the indicated helicities being opposite. It appears to be a characteristic feature of the twistor theory that particles are represented by *pairs* of twistor scalar product lines. In the case of a free photon entering and leaving without interaction, we can represent the scalar product diagram in either of the ways indicated in fig. 27. Observe that when the two central lines are separated, the helicities proceed in the same direction, rather than in the opposite direction as appears to be the case for a virtual photon exchange. In the photon exchange, the two segments bearing "−1" may, perhaps, be thought of as the "same" type of segment as those into which the free photon appears to split in the right-hand diagram of fig. 27. The difference is that in the photon exchange, the lines have become quite disconnected, with the helicities working in opposition rather than together. Thus, the virtual photon, as a whole, does not have (as it should not have) a well-defined helicity.

In this theory we may think of the twistor situation depicted in fig. 28 as playing the part of the Feynman vertex of the conventional theory. However, here the photon line is split into two separate parts which cannot by themselves be assembled to yield a single external photon line. This seems to reflect the fact that a single Feynman vertex represents a virtual process. But how, then, are we to represent real processes built up from several Feynman vertices in which external photon lines may be present? Let us consider the simplest case, namely Compton scattering. We may attempt to make the correspondence indicated in fig. 29. The external photon lines are not yet assembled correctly, however, since we need to produce double lines numbered "+2" to re-

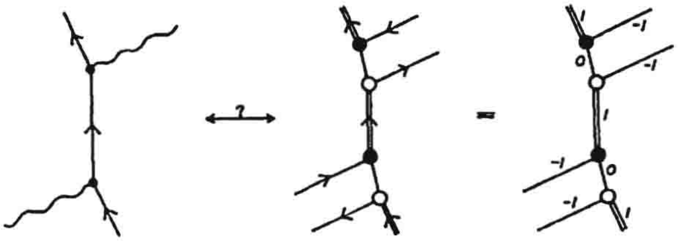


Fig. 29.

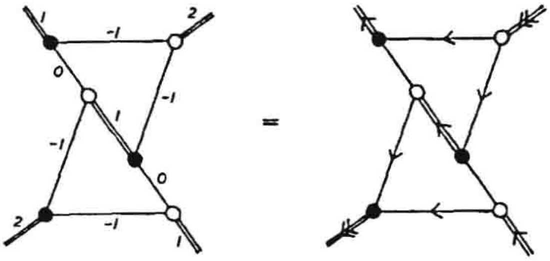


Fig. 30.

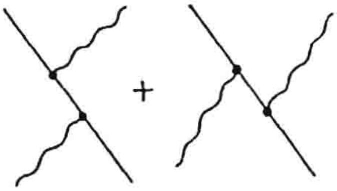
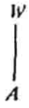


Fig. 31.

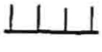
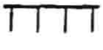
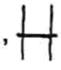
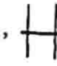
present external photons. It becomes clear that to achieve this according to the rules for the twistor diagrams we must assemble each external photon using one twistor line from *each* "Feynman vertex". The result is given in fig. 30. Note that the two photons attach themselves to the electron in a way which does not allow one to say that one interaction occurred completely "before" the other one. Thus, instead of simply describing the Feynman diagram in fig. 29, we must think of fig. 30 as describing the complete process indicated in fig. 31, except that we have singled out a particular helicity state for the photon as well as for the (massless) electron. (Of course, in the conventional formalism, these two diagrams must be added in any case, in order to obtain a meaningful answer.) Remarkably enough, computation of the twistor integral of fig. 30 does in fact yield the correct angular dependence for Compton scattering of massless (i.e. high energy limit) electrons.

4.3. *Evaluation of diagrams using elementary states*

So far we have not attempted to *compute* any of the twistor integrals. Not surprisingly these computations can be difficult to perform. The integrals are taken over manifolds (which are themselves not always easy to locate) of rather large dimension. Also, twistor diagrams involve "closed loops" even for the simplest processes where the conventional formalism yields "trees". It often turns out that indirect methods of evaluating the integrals must be used if the computations are not to get out of hand. But let us just start by evaluating a few very simple twistor diagrams, without regard to the question of physical meaningfulness. The first diagram we shall evaluate is that of fig. 32. However, before doing so we shall introduce some pictorial notation which will help us keep track of the contractions and skewing of the twistors involved in the integrals (cf. [1]). This is as follows:



stands for $W_\alpha A^\alpha$

 stands for $\epsilon^{\alpha\beta\gamma\delta}$,  for $\epsilon_{\alpha\beta\gamma\delta}$,  for $2\delta_\rho^{[\alpha}\delta_\sigma^{\beta]}$,  for $6\delta_\rho^{[\alpha}\delta_\sigma^{\beta}\delta_\tau^{\gamma]}$,

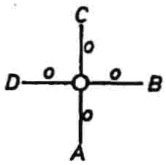
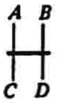
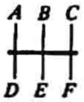
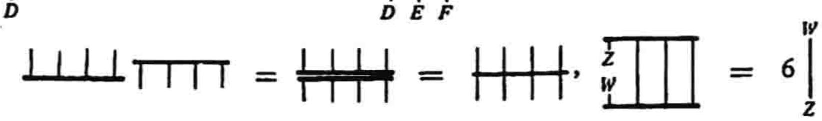
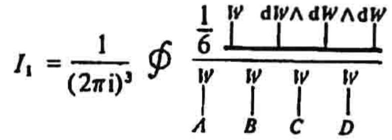


Fig. 32.

so  stands for $2A_\alpha B_\beta C^{[\alpha}D^{\beta]}$,  for $6A_\alpha B_\beta C_\gamma D^{[\alpha}E^\beta F^{\gamma]}$, etc.

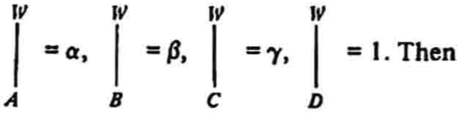
Note that  etc.

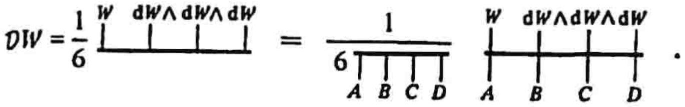
As an example consider fig. 32. The relevant contour integral is

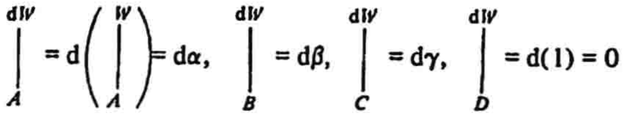

$I_1 = \frac{1}{(2\pi i)^3}\oint$
1/6
W dW∧dW∧dW
W W W W
A B C D

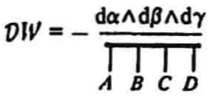
(4.6)

the contour being $S^1 \times S^1 \times S^1$.

Put 
W
A
= α,
W
B
= β,
W
C
= γ,
W
D
= 1. Then


𝒟W = 1/6
W dW∧dW∧dW
=
1
6
A B C D
W dW∧dW∧dW
A B C D

Now 
dW
A
= d
W
A
= dα,
dW
B
= dβ,
dW
C
= dγ,
dW
D
= d(1) = 0


𝒟W = −
dα∧dβ∧dγ
A B C D

(4.7)

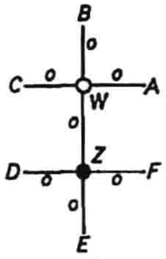

B
o
C
o
o
A
W
o
Z
D
o
o
F
o
E

Fig. 33.

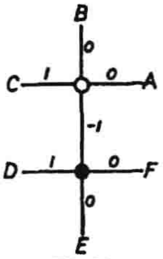
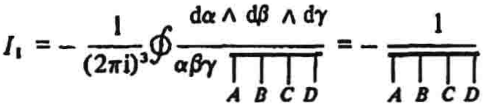
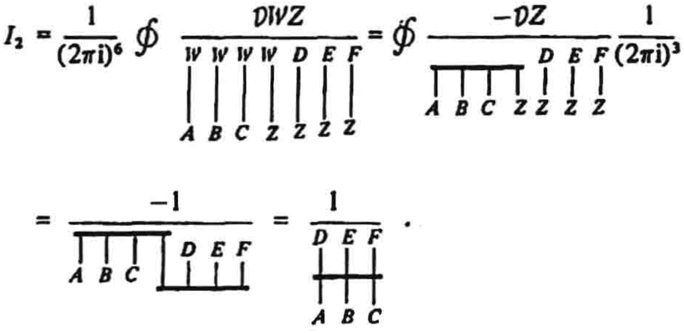

B
o
C
1
o
A
-1
D
1
o
F
o
E

Fig. 34.


$I_1 = -\frac{1}{(2\pi i)^3}\oint \frac{d\alpha\wedge d\beta\wedge d\gamma}{\alpha\beta\gamma}$
A B C D
$= -\frac{1}{}$
A B C D

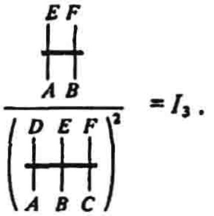
(4.8)

Similarly we may evaluate fig. 33 which gives


$I_2 = \frac{1}{(2\pi i)^6}\oint$
DWZ
W W W W D E F
A B C Z Z Z Z
$= \oint$
−DZ
D E F
$\frac{1}{(2\pi i)^3}$
A B C Z Z Z Z
= −1
A B C
D E F
= 1
D E F
A B C

(4.9)

Now we remark that the value of fig. 34 may be found from that of fig. 33 by differentiating the value of fig. 33 with respect to C^α and selecting the coefficient of D_α in the answer (as can be seen from an integration by parts of the integrand of fig. 34). The value of fig. 34 is thus


E F
A B
D E F
2
A B C
$= I_3$.

(4.10)

By combining these processes we can evaluate any tree diagram with single external lines, which has non-negative integers on its external lines.

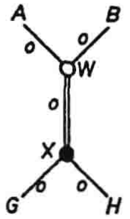


Fig. 35.

For fig. 35 we have

$$I_4 = \frac{1}{(2\pi i)^5}\oint \frac{\mathcal{D}WX}{\begin{array}{c}W\\|\\A\end{array}\ \begin{array}{c}W\\|\\B\end{array}\left(\begin{array}{c}W\\|\\X\end{array}\right)^2 \begin{array}{c}G\\|\\X\end{array}\ \begin{array}{c}H\\|\\X\end{array}} \tag{4.11}$$

where the contour integration is made up of four S^1 integrals and one S^2 integral taken in succession. (Alternatively the surface of integration can involve an S^3 explicitly, this being a twisted product of S^1 with S^2.) Assume that in the **C**-picture the line of intersection of the two planes, G, H does not meet the line AB (otherwise I_4 becomes infinite). Choose auxiliary twistors P^α, Q^α, R_α, S_α, where for simplicity, R and S are general planes through AB, and $P = G \cap H \cap S$, $Q = G \cap H \cap R$. Then, normalizing appropriately, we have

$$\begin{array}{c}R\\|\\A\end{array} = \begin{array}{c}R\\|\\B\end{array} = \begin{array}{c}R\\|\\Q\end{array} = \begin{array}{c}S\\|\\A\end{array} = \begin{array}{c}S\\|\\B\end{array} = \begin{array}{c}S\\|\\P\end{array} = \begin{array}{c}G\\|\\P\end{array} = \begin{array}{c}H\\|\\P\end{array} = \begin{array}{c}G\\|\\Q\end{array} = \begin{array}{c}H\\|\\Q\end{array} = 0, \qquad \begin{array}{c}R\\|\\P\end{array} = \begin{array}{c}S\\|\\Q\end{array} = 1. \tag{4.12}$$

Put

$$\begin{array}{c}W\\|\\A\end{array} = \alpha,\ \begin{array}{c}W\\|\\B\end{array} = \beta,\ \begin{array}{c}W\\|\\P\end{array} = \psi,\ \begin{array}{c}W\\|\\Q\end{array} = \chi,\ \begin{array}{c}G\\|\\X\end{array} = \gamma,\ \begin{array}{c}H\\|\\X\end{array} = \eta,\ \begin{array}{c}R\\|\\X\end{array} = \rho,\ \begin{array}{c}S\\|\\X\end{array} = \sigma.$$

The S^2 integration involves the variables ψ, χ, ρ, σ given by

$$\psi = e^{i\varphi}\cos\tfrac{1}{2}\theta, \qquad \chi = \sin\tfrac{1}{2}\theta, \qquad \rho = e^{-i\varphi}\cos\tfrac{1}{2}\theta, \qquad \sigma = \sin\tfrac{1}{2}\theta.$$

Since ψ, χ, ρ, σ involve only two independent variables, the wedge product of more than two of them must vanish. It follows that

$$\mathcal{D}WX = \left(\begin{array}{cccc} G & H & R & S\\ \hline | & | & | & |\\ A & B & P & Q\end{array}\right)^{-1} d\alpha \wedge d\beta \wedge (\psi d\chi - \chi d\psi) \wedge d\gamma \wedge d\eta \wedge (\rho d\sigma - \sigma d\rho)$$

whence

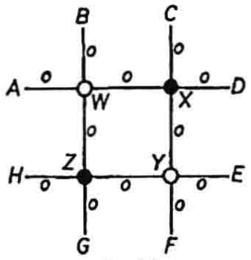
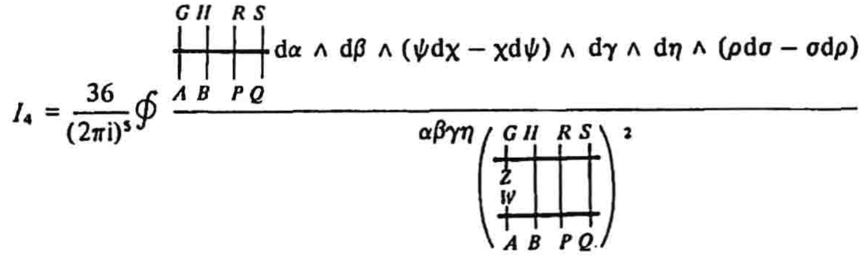
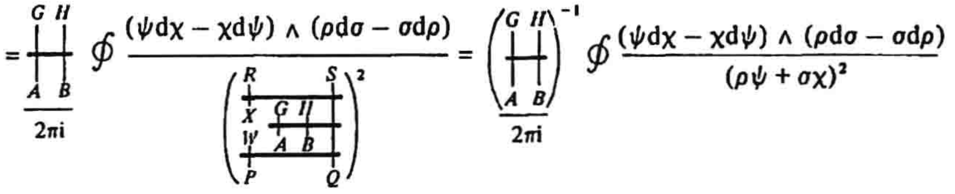
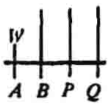
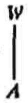
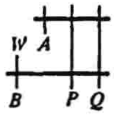
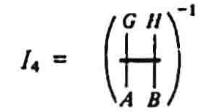


Fig. 36.



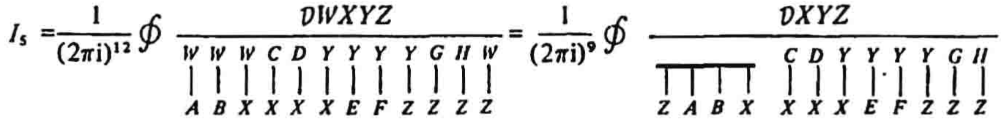


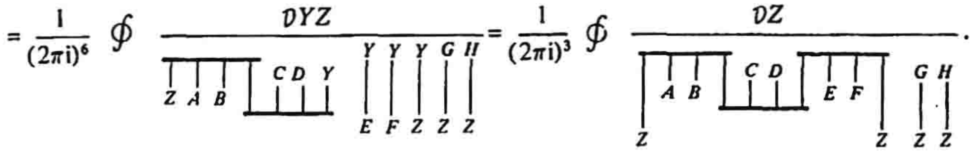
since the part of  independent of $\alpha =$  is $-\frac{1}{2}$  , etc., and where (4.12) has been made use of. Finally, integrating over the spherical polar variables θ and φ, we obtain



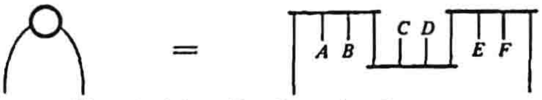
(4.13)

Fig. 36 provides a somewhat more complicated example. We have




$= \frac{1}{(2\pi i)^6} \oint$
$\mathcal{D}YZ$
Z A B
C D Y
Y Y Y G H
E F Z Z Z
$= \frac{1}{(2\pi i)^3} \oint$
$\mathcal{D}Z$
A B
C D
E F
G H
Z
Z Z Z

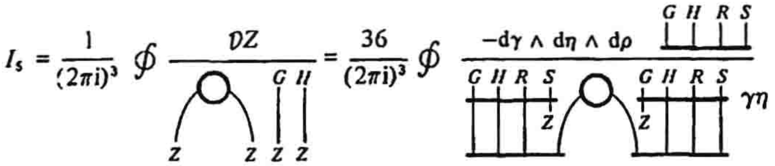
Write

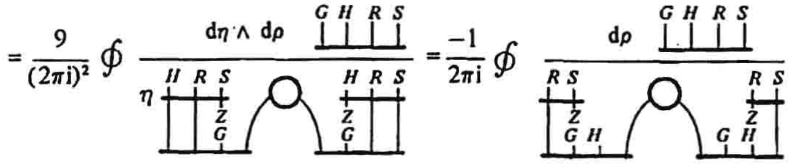


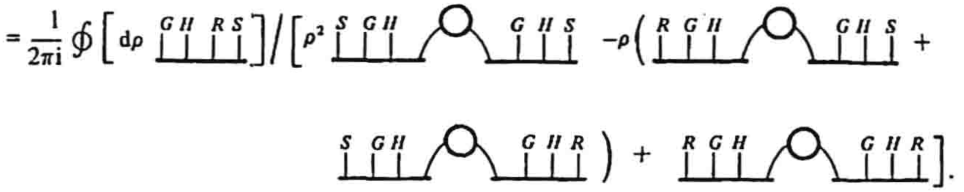
Introduce auxiliary twistors R_α, S_α and put

$$\overset{G}{\underset{Z}{|}} = \gamma, \quad \overset{H}{\underset{Z}{|}} = \eta, \quad \overset{R}{\underset{Z}{|}} = \rho, \quad \overset{S}{\underset{Z}{|}} = 1.$$

Then


$I_s = \frac{1}{(2\pi i)^3} \oint$
$\mathcal{D}Z$
G H
Z
Z Z Z
$= \frac{36}{(2\pi i)^3} \oint$
$-d\gamma \wedge d\eta \wedge d\rho$
G H R S
G H R S
Z
G H R S
Z
$\gamma\eta$


$= \frac{9}{(2\pi i)^2} \oint$
$d\eta \wedge d\rho$
G H R S
η
H R S
Z
G
H R S
Z
G
$= \frac{-1}{2\pi i} \oint$
$d\rho$
G H R S
R S
Z
G H
R S
Z
G H


$= \frac{1}{2\pi i} \oint$
$d\rho$
G H R S
ρ^2
S G H
G H S
$-\rho$
R G H
G H S
+
S G H
G H R
+
R G H
G H R

Taking the contour to pass between the two zeros of the quadratic expression in the denominator we obtain, after some manipulation,

$$I_s = (2ab + 2bc + 2ca - a^2 - b^2 - c^2)^{1/2}$$

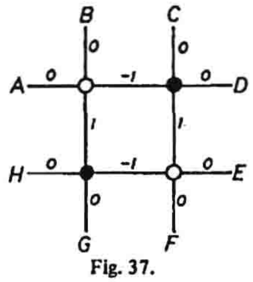


Fig. 37.

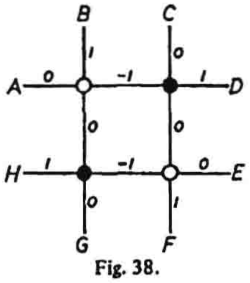
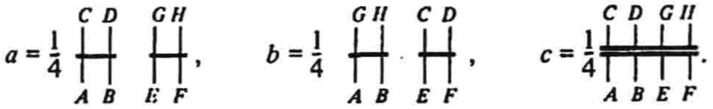


Fig. 38.

where



It can be shown that fig. 37 leads to a value

$$I_6 = \left(1 + \frac{a-c}{b}\right) I_5 .$$

Then, one has for fig. 38 (high energy limit of Möller scattering in terms of elementary states)

$$I_7 = \frac{\partial^2}{\partial D^\nu \partial F_\nu} \cdot \frac{\partial^2}{\partial B_\mu \partial H^\mu} I_6 . \qquad (4.17)$$

4.4. *Momentum state diagrams*

Finally we want to consider states, which are momentum eigenstates, represented by

$$f_n(Z) = \frac{\exp(K\cdot Z/B\cdot Z)}{(A\cdot Z)(B\cdot Z)^{n+1}} \qquad (4.18)$$

where A and B are generators of the null cone at infinity and K a null line meeting A. Momentum states are needed in order to obtain direct comparisons with the usual results of quantum mechanics, since elementary states are difficult to interpret. (4.18) represents a circularly polarised plane wave in the direction of A. We have[28]

$$\begin{aligned} \phi_0(Y, Z) &= \frac{1}{2\pi i} \oint f(\lambda Z + Y) \mathrm{d}\lambda \\ &= \frac{(A\cdot Z)^n \exp(AK{:}P/AB{:}P)}{(AB{:}P)^{n+1}} \end{aligned} \qquad (4.19)$$

---

[28] Where $AB{:}P = A_\alpha B_\beta P^{\alpha\beta}$.

and $P_{\alpha\beta}$ (i.e. $x^{QQ'}$) represents the point of intersection of Y and Z.[29] We have $I_{\alpha\beta} = A_\alpha B_\beta - A_\beta B_\alpha$ and if $A_\alpha = (0, \kappa^{A'})$, $B_\alpha = (0, \tau^{A'})$, $K_\alpha = (\bar{\kappa}_A, 0)$ and $\epsilon^{A'B'} = \kappa^{A'}\tau^{B'} - \kappa^{B'}\tau^{A'}$ $(\kappa_{A'}\tau^{A'} = 1)$. The field is null, with principal null direction that of A, as $A\cdot Z$ gives a simple pole. $AK{:}P = -\mathrm{i}\bar{\kappa}_Q\kappa_{Q'}x^{QQ'} = -\mathrm{i}k_a x^a$ and

$$\phi_0(Y, Z) = (\kappa^{A'}\pi_{A'}) \exp(-\mathrm{i}k^a x_a)$$

$$\phi^{A'B'\ldots L'} = \kappa^{A'} \ldots \kappa^{L'} \exp(-\mathrm{i}k_q x^q).$$

(For Maxwell fields this expression was given in section 1.2.)

Now we wish to evaluate figs. 17 and 30 using momentum states. In this case the contours will no longer be well-defined as they become pinched. We expect, and find, a δ-function singularity. To calculate the scatterings given by figs. 17 and 30, i.e. to find the coefficient of the δ-function, we can proceed formally by making use of the operators

i) $I^{\alpha\beta}\,\partial/\partial Z^\beta$. This acts on f_n to give

$$I^{\alpha\beta}\frac{\partial}{\partial Z^\beta} f_n = \frac{I^{\alpha\beta}K_\beta}{(B\cdot Z)}\,\frac{\exp(K\cdot Z/B\cdot Z)}{(A\cdot Z)(B\cdot Z)^{n+1}}$$

$$= \frac{\bar{A}^\alpha}{(B\cdot Z)}\,\frac{\exp(K\cdot Z/B\cdot Z)}{(A\cdot Z)(B\cdot Z)^{n+1}} = \bar{A}^\alpha f_{n+1} \tag{4.20}$$

since $I^{\alpha\beta}A_\beta = I^{\alpha\beta}B_\beta = 0$.

ii) $I_{\alpha\beta}Z^\beta$. This yields

$$I_{\alpha\beta}Z^\beta f_n = \{A_\alpha(B_\beta Z^\beta) - B_\alpha(A_\beta Z^\beta)\}\, f_n$$

$$= A_\alpha(B_\beta Z^\beta) f_n \qquad \text{in integrals,}$$

$$= A_\alpha f_{n-1}\,. \tag{4.21}$$

This is because $\oint f_n(A\cdot Z)\mathcal{D}Z = 0$, since the contour separates the singularities of A and B and multiplication by $(A\cdot Z)$ enables us to contract the contour through what was formerly the singularity due to $(A\cdot Z)^{-1}$.

iii) $Z^\alpha I^{\beta\gamma}\partial f_n/\partial Z^\gamma$. This is represented by

$$\left(Z^\alpha I^{\beta\gamma}\frac{\partial}{\partial Z^\gamma} + Y^\alpha I^{\beta\gamma}\frac{\partial}{\partial Y^\gamma}\right)\phi_0(Y, Z) = A_\lambda P^{\lambda\alpha}\bar{A}^\beta\phi_0 \tag{4.22}$$

$$\leftrightarrow \left(\begin{array}{c|c} -\mathrm{i}\kappa_{C'}x^{AC'}\bar{\kappa}^B & \kappa_{A'}\bar{\kappa}^B \\ \hline 0 & 0 \end{array}\right)\phi_0 = \left(\begin{array}{c|c} \bar{\kappa}^B\,\partial\phi_0/\partial\bar{\kappa}_A & \bar{\kappa}^B\kappa_{A'}\phi_0 \\ \hline 0 & 0 \end{array}\right). \tag{4.23}$$

Applying operator iii) and its dual version to two connected vertices and contracting over α, we find the effect is to multiply the numerator of the integrand by $(W\cdot Z)$ changing $(W\cdot Z)_a$ to $-a(W\cdot Z)_{a-1}$.

Now we want to evaluate the cross-sections arising from diagrams of the form of fig. 39a, which we abbreviate to fig. 39b. Here the states are labelled by the spinors $\kappa^{A'}$, ρ^A, $\sigma^{A'}$, τ^A corresponding to the momentum vectors:

---

[29] That ϕ_0, the null-datum on the null cone at infinity, represents the full field, is proved in [11, 33]; see also the discussion in [21] concerning such waves and their representation.

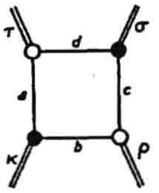


Fig. 39a.

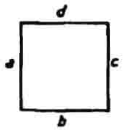
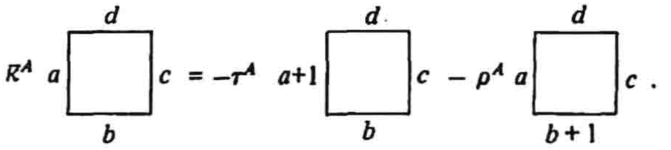


Fig. 39b.

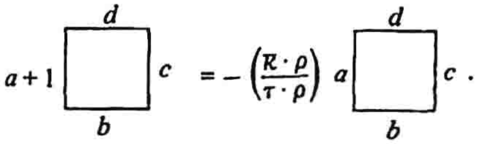
$$k^a = \bar{\kappa}^A \kappa^{A'}, \qquad r^a = \rho^A \bar{\rho}^{A'}, \qquad s^a = \bar{\sigma}^A \sigma^{A'}, \qquad t^a = \tau^A \bar{\tau}^{A'}.$$

Using operator i) and integrating by parts we find



(4.24)

Hence


a+1
d
c
b
$= -\left(\frac{\bar{\kappa}\cdot\rho}{\tau\cdot\rho}\right)$ a
d
c .
b

(4.25)

Using this formula we find we need only evaluate one of the box diagrams.

Now, let us consider (say) the case when k^a and s^a are incoming and define

$$\delta = \delta(\kappa\bar{\kappa} - \rho\bar{\rho} + \sigma\bar{\sigma} - \tau\bar{\tau}) = \delta(k^a - r^a + s^a - t^a).$$

Define the relation $\approx$ by

$$\theta \approx \psi \Leftrightarrow \theta\delta = \psi\delta,$$

i.e. "$\theta \approx \psi$" means $\theta - \psi$ vanishes whenever $k^a + s^a = r^a + t^a$. Then

$$\alpha{:}= -\bar{\kappa}\cdot\rho/\tau\cdot\rho \approx -\bar{\tau}\cdot\sigma/\kappa\cdot\sigma =: \alpha'$$

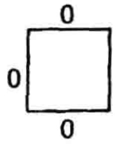
$$\beta{:}= -\bar{\sigma}\cdot\tau/\rho\cdot\tau \approx -\bar{\rho}\cdot\kappa/\sigma\cdot\kappa =: \beta'$$

$$\lambda{:}= -\bar{\kappa}\cdot\tau/\rho\cdot\tau \approx -\bar{\rho}\cdot\sigma/\kappa\cdot\sigma =: \lambda'$$

$$\mu{:}= -\bar{\sigma}\cdot\rho/\tau\cdot\rho \approx -\bar{\tau}\cdot\kappa/\sigma\cdot\kappa =: \mu'.$$

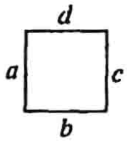
The variables $\alpha, \beta, \lambda, \mu$ are related to the Mandelstam variables s, t, u used in quantum mechanical scattering calculations. The extra information conveyed by the extra variable relates to the phase (in its connection with the spin).

Now *if*



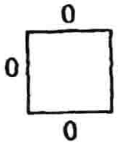
$= \delta$

then



$= \alpha^a\beta^b\lambda^c\mu^d\delta.$ (4.26)

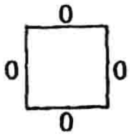
To prove the supposition here, we apply (4.22, 4.23) to



$= \chi$ and find

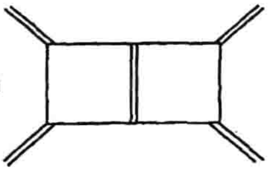
$$\tau_A\,\partial\chi/\partial\bar\kappa_A + \kappa_{A'}\,\partial\chi/\partial\bar\tau_{A'} = 0, \text{ etc.} \tag{4.27}$$

Hence

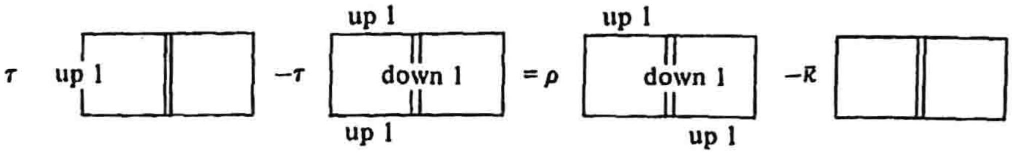


is a function of $(\kappa\bar\kappa - \rho\bar\rho + \sigma\bar\sigma - \tau\bar\tau)$ alone and it must be the δ-function to ensure conservation of momentum.

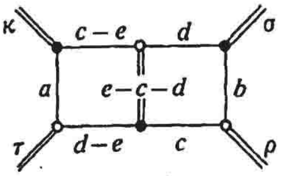
To cope with diagrams of the form



is not so simple. We can set up a series of relations such as



and differential operator relations analogous to (4.20)–(4.23) and eventually eliminate to derive a (soluble) differential equation. The result is



$$= \alpha^{a-e}\beta^{b}\lambda^{c}\mu^{d}\left\{\sum_r \binom{a}{r+e}\binom{b}{r}\left(\frac{\lambda\mu}{\alpha\beta}\right)^{r}\right\}\delta \qquad (4.28)$$

$$= \alpha^{a-e}\beta^{b}\lambda^{c}\mu^{d}\left(1-\frac{\lambda\mu}{\alpha\beta}\right)^{b+a+1}\left\{\sum_r \binom{a+r}{r+e}\binom{b+r+e}{r}\left(\frac{\lambda\mu}{\alpha\beta}\right)^{r}\right\}\delta$$

where $\binom{a}{b}$ is the usual binomial coefficient.

To find the differential cross-sections from these amplitudes we must take the squares of their moduli. We may verify that

$$\bar{\alpha} = \frac{\beta'}{\alpha'\beta' - \lambda'\mu'}\,, \qquad \bar{\lambda} = \frac{\mu'}{\lambda'\mu' - \alpha'\beta'}\,,$$

so

$$|\alpha|^2 \approx |\beta|^2 \approx 1 - |\lambda|^2 \approx 1 - |\mu|^2 .$$

The quantities

$$|\alpha|^2 = k_a r^a / t_b r^b\,, \qquad |\lambda|^2 = k_a t^a / r_b t^b\,, \text{ etc.}$$

can be expressed directly in terms of the scattering angle. Explicitly we have (taking the momentum conservation relation to hold)

$$|\alpha| = |\beta| = C, \qquad |\lambda| = |\mu| = S$$

provided k^a and s^a are incoming (with $k^a \to t^a$, $s^a \to r^a$), or else if t^a and r^a are incoming ($t^a \to k^a$, $r^a \to s^a$), where we have set

$$C = \cos\tfrac{1}{2}\theta, \qquad S = \sin\tfrac{1}{2}\theta,$$

the scattering angle θ being measured in the centre-of-mass frame.

Taking account of the modifications that are required when *other* pairs of states are the incoming ones, we have:

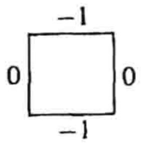
$$|\alpha| = |\beta| = 1/C, \qquad |\lambda| = |\mu| = S/C$$

if k^a and r^a are incoming ($k^a \to t^a$, $r^a \to s^a$) or else s^a and t^a are incoming ($s^a \to r^a$, $t^a \to k^a$); or

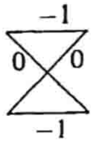
$$|\alpha| = |\beta| = C/S, \qquad |\lambda| = |\mu| = 1/S$$

if k^a and t^a are incoming ($k^a \to s^a$, $t^a \to r^a$), or else r^a and s^a are incoming ($r^a \to t^a$, $s^a \to k^a$). (Such manifestations of crossing symmetry must arise automatically in this theory owing to the analyticity of the operations involved.)

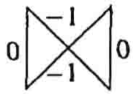
Let us consider the high energy limit of Möller scattering (fig. 11). We represent time as proceeding up the page in our diagrams. Thus we have



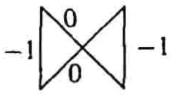
and



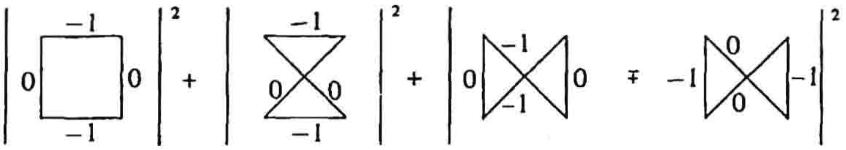
when the helicities of the two incoming (massless) electrons are opposite and



and



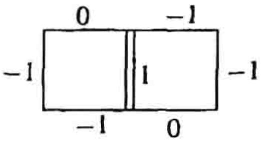
if the helicities are the same. Because of Fermi–Dirac statistics applying when the helicities are the same we must combine these last two amplitudes coherently, i.e. subtract them from one another – or else add them (since a sign uncertainty arises here owing to the lack of a definite rule, as yet, for choosing the contour orientations) – before taking the squares of their moduli. Summing (incoherently) over the different helicities we get, for unpolarized electrons



$$= \left\{\left(\frac{C^2}{S^2}\right)^2 + \left(\frac{S^2}{C^2}\right)^2 + \left(\frac{1}{S^2} + \frac{1}{C^2}\right)^2\right\}\delta$$

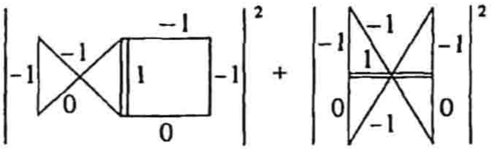
for the differential cross-section (ignoring possible overall factors and taking the plus sign so as to agree with the correct answer [48]!).

The calculation for Compton scattering (or equivalently, by crossing symmetry, for pair annihilation) is similar, where we use



$$= \mu^{-1}(\alpha\beta - \lambda\mu)^{-1}\,\delta.$$

The comparison for pair annihilation is as follows:



$$= \left\{\left(\frac{S}{C}\right)^2 + \left(\frac{C}{S}\right)^2\right\}\delta$$

in agreement (ignoring an overall factor) with the standard result [48].

5. Further developments

In this section some of the later developments in twistor theory, which have evolved since the lecture course in 1970, will be outlined. These developments are still in progress and it will not be appropriate to go into them in full detail, but the main ideas will be discussed.

5.1. *Asymptotic twistors*

Recall the definition of a null global twistor Z^α in curved space-time, given in section 3.2, as a null geodesic Z along which a spinor $\pi_{A'}$ is defined, the flagpole direction of $\pi_{A'}$ being everywhere tangent to the null geodesic and $\pi_{A'}$ being propagated parallely along it. The basic difficulty with this is that we have too little structure; all we obtain is a symplectic structure, and non-null twistors are not even defined. There is no complex structure, essential if contour integrals are to have a meaning; and the basic ingredient of the contour integrals we have been considering, namely the twistor scalar product, is also missing. However, if our intention is to employ such concepts in the construction of an S-matrix theory applicable to curved space-times, the situation is considerably more encouraging. For then we can reasonably restrict attention to space-times which are, in the appropriate sense, asymptotically flat. We shall assume, also, that the Einstein vacuum or Einstein-Maxwell equations (without cosmological term) hold at large distances. Then each null geodesic starts out from a region, remote from massive sources, where the space-time is flat enough that additional twistor structure of the type required can be assigned – and likewise it ends up in such a region. It turns out that the normal assumptions of asymptotic flatness are just sufficient that non-null twistors, a complex structure, and a form of scalar product can all be defined. However, enough residual space-time curvature is still present at infinity to give a twistor space structure essentially *different* from that of Minkowski space-time. In effect, the twistor space becomes *curved* (as a Kähler manifold).

To carry out a detailed discussion, we shall require the picture of conformal infinity appropriate to an asymptotically flat space-time $\mathcal{M}$. For details, see [2, 4, 41]. We have already considered compactified Minkowski space in section 1.1. The essential new feature in the case of an asymptotically flat space is that past null infinity and future null infinity cannot be usefully identified with each other. By introducing a convenient conformal factor which falls off suitably towards null infinity, we obtain a new metric $\hat{g}_{ab} = \Omega^2 g_{ab}$ which remains finite and smooth along the two adjoined bounding null hypersurfaces $\mathscr{I}^-$, representing past null infinity, and $\mathscr{I}^+$, representing future null infinity. The picture is that of fig. 1 but without the identification.

In a scattering problem, the data specifying the in-states (assumed zero rest-mass) would be given on $\mathscr{I}^-$ and those specifying the out-states on $\mathscr{I}^+$. Each point of $\mathscr{I}^-$ is the vertex of a light cone whose generators, namely the null geodesics through that point, would describe a null hypersurface in $\mathcal{M}$ which is asymptotically plane in the past. This hypersurface could therefore be thought of as a constant phase hypersurface of an incoming asymptotic plane wave. (Ignore dispersion effects.) Similarly, each point of $\mathscr{I}^+$ may be associated with a constant phase hypersurface of an outgoing asymptotic plane wave.

Every null geodesic in $\mathcal{M}$ acquires a past end-point on $\mathscr{I}^-$ and a future end-point on $\mathscr{I}^+$ (taking "asymptotic simplicity" [2, 4], as a necessary condition for asymptotic flatness); also there is a null geodesic *on* $\mathscr{I}^\pm$ through each point on $\mathscr{I}^\pm$ – since $\mathscr{I}^\pm$, being null hypersurfaces, are generated by null geodesics. The different points of one generator of $\mathscr{I}^-$ represent the different constant phase hypersurfaces of the *same* incoming asymptotic plane wave; similarly, the points of a generator of $\mathscr{I}^{+}$ represent the various phase hypersurfaces of one outgoing asymptotic plane wave. Each generator of $\mathscr{I}^{-}$ can be shown to be naturally associated with a corresponding "opposite" generator of $\mathscr{I}^{+}$. That is, each incoming null direction defines a unique "undeflected" outgoing null direction. For a *given* null geodesic in $\mathcal{M}$, the incoming and outgoing directions will in general be different (i.e. "deflection" occurs). There appears to be no natural association of an individual *point* of $\mathscr{I}^{-}$ with some uniquely corresponding *point* of $\mathscr{I}^{+}$. This means that we have no natural meaning for the phase shift between ingoing and outgoing asymptotically plane waves even when their directions agree.

Not only is the relation between $\mathscr{I}^{-}$ and $\mathscr{I}^{+}$ different, for an asymptotically flat space, from the situation for Minkowski space, but the structure of each of $\mathscr{I}^{-}$ and $\mathscr{I}^{+}$ contains information about the asymptotic gravitational field. It turns out that the Weyl tensor C_{abcd} or equivalently the spinor[30] Ψ_{ABCD} must vanish on $\mathscr{I}^{\pm}$. But the quantity $K_{abcd} = \Omega^{-1} C_{abcd}$ (equivalently $\phi_{ABCD} = \Omega^{-1}\Psi_{ABCD}$) does not (cf. (1.27)) in general. The component $\Psi_4 := \phi_{ABCD}\iota^A\iota^B\iota^C\iota^D$ in fact measures the gravitational radiation field, incoming if on $\mathscr{I}^{-}$ and outgoing if on $\mathscr{I}^{+}$, where ι^A points along the generators of $\mathscr{I}^{\pm}$. Ψ_4^0 is the derivative along generators of $\mathscr{I}^{\pm}$ of the Bondi–Sachs news function N, whose squared modulus $N\overline{N}$ may be taken to represent the energy-flux of gravitational radiation across $\mathscr{I}^{\pm}$. The quantity N governs the rate of change, along the generators of $\mathscr{I}^{\pm}$, of the asymptotic shear σ^0 of (Bondi-type) systems of advanced or retarded time hypersurfaces.

We shall next define the space C of asymptotic twistors. In fact there are two such spaces depending upon whether we use $\mathscr{I}^{-}$ or $\mathscr{I}^{+}$ on which to define things. For definiteness, let us choose $\mathscr{I}^{+}$, but we must bear in mind that a parallel discussion in terms of $\mathscr{I}^{-}$ can also be given. (For a *retarded* gravitational field we would expect to obtain a less interesting structure for $\mathscr{I}^{-}$ than for $\mathscr{I}^{+}$, however.) Consider local twistors $(\omega^A, \pi_{A'})$ (cf. section 3.1) defined at points of $\mathscr{I}^{+}$. We might think that since the conformal (Weyl) curvature vanishes at every point of $\mathscr{I}^{+}$, we could use local twistor transport, i.e.

$$t^b\nabla_b\omega^A = -\mathrm{i}\, t^{AB'}\pi_{B'}, \qquad t^b\nabla_b\pi_{A'} = -\mathrm{i}\, P_{BCB'A'}t^{BB'}\omega^C \tag{5.1}$$

(cf. (3.8)), to carry the local twistor $(\omega^A, \pi_{A'})$ along a curve on $\mathscr{I}^{+}$ with tangent vector t^b and hence obtain an *integrable* equivalence between local twistors at different points on $\mathscr{I}^{+}$. This would lead to a single twistor space for the whole of $\mathscr{I}^{+}$. Unfortunately, things are not as simple as this since the commutator of local twistor derivatives involves the *derivative* of the Weyl conformal spinor (in fact the quantity $\nabla^A_{A'}\Psi_{ABCD}$, cf. (3.11)) and this does *not* vanish on $\mathscr{I}^{+}$, in general. Thus, if $(\omega^A, \pi_{A'})$ is carried by local twistor transport from one point of $\mathscr{I}^{+}$ to another over two different routes each lying on $\mathscr{I}^{+}$, then we must expect that in general unequal results will be obtained.

We must therefore proceed somewhat differently. Let us consider local twistors on $\mathscr{I}^{+}$ which have a spinor representation of the form

$$(\xi\iota^A, \pi_{A'}) \tag{5.2}$$

where a particular selection of the spinor ι^A at each point (with flagpole pointing along the generator of $\mathscr{I}^+$ through the point) is made. We can use local twistor transport along generators of $\mathscr{I}^+$ to equate local twistors of this form at different points on the *same* generator. (The form (5.2) is preserved by (5.1), if we propagate along generators ($t^a = \iota^A \iota^{A'}$); also no problems from non-integrability can arise since the generators are one-dimensional simply-connected lines.) Thus, associated with every generator of $\mathscr{I}^+$ we have a 3-complex-dimensional vector space, parameterized at any one point of the generator by ξ and the two components of $\pi_{A'}$. The generators of $\mathscr{I}^+$ themselves can be labelled by one further complex parameter ζ. Thus we can assemble a 4-complex-dimensional manifold of quantities of this kind. This is the "blown-up" manifold of (future) asymptotic twistors.

---

[30] Strictly speaking, these should be denoted $\hat{C}_{abcd}$ and $\hat{\Psi}_{ABCD}$, respectively, since they refer to the metric $\hat{g}_{ab}$ and not g_{ab}. However, we shall, for convenience, drop all "hats" in this section, it being assumed that the metric $\hat{g}_{ab}$ is being always referred to.

We shall give some structure to this manifold shortly. For the moment, let us explain how this picture arises, a little more fully. Consider an ordinary global null twistor Z^α (cf. (3.2)) in $\mathbb{M}$. The null geodesic Z which it defines intersects $\mathscr{I}^+$ at a single point P, say. We have $\pi_{A'}$ parallelly propagated along Z with respect to both the physical and conformally rescaled metrics (as follows from (1.19a) since $\bar{\pi}^A \pi^{A'} \nabla_{AA'}$ defines propagation along Z). Thus we can define $\pi_{A'}$ at P as well as at finite points of Z. Furthermore, knowledge of the point P on $\mathscr{I}^+$ and of $\pi_{A'}$ at P will be sufficient to determine the twistor Z^α. Let us label Z^α by the local twistor $(0, \pi_{A'})$ at P.

This is of the form (5.2) at P, so local twistor transport along the generator γ of $\mathscr{I}^+$ through P, to a point Q, yields another local twistor of the form $(\xi\iota^A, \pi_{A'})$. Knowing this local twistor at Q we can locate P simply by finding the point at which ξ is reduced to zero. The $\pi_{A'}$ at that point enables us to locate the global twistor Z. Of course not *every* $(\xi\iota^A, \pi_{A'})$ at Q yields a point P at which ξ is reduced to zero, since ξ is *complex* and we have one *real* dimension of points on γ. The condition that P does exist, assuming that $\pi_{A'}$ is not proportional to $\iota_{A'}$, is that the local twistor $(\xi\iota^A, \pi_{A'})$ be *null*, i.e.

$$\mathrm{Re}(\xi\pi_{1'}) = 0 \tag{5.3}$$

where $\pi_{1'} = \pi_{A'}\iota^{A'}$. Note that $Z^\alpha \bar{Z}_\alpha$ is preserved under local twistor transport.

If u is a parameter asssociated with the tangent vector $\iota^A\iota^{A'}$ to γ (i.e. $\nabla_{11'} := \iota^A\iota^{A'}\nabla_{AA'} = \mathrm{d}/\mathrm{d}u$ when applied to scalar functions on γ) and if we choose ι^A to be parallelly propagated along γ (i.e. $\nabla_{11'}\iota^B = 0$) then (5.1) with $t^a = \iota^A \iota^{A'}$ gives

$$\mathrm{d}\xi/\mathrm{d}u = -\mathrm{i}\pi_{1'}. \tag{5.4}$$

It is convenient to choose the conformal factor Ω so that the divergence of the generators of $\mathscr{I}^+$ vanishes. The shear of $\mathscr{I}^+$ necessarily also vanishes [4]. The generators of $\mathscr{I}^+$ then become parallel to one another, in the sense that every cross-section of $\mathscr{I}^+$ (a spacelike 2-surface) has the same metric (induced by $\hat{g}_{ab}$) as every other. Under these circumstances it can be shown [4, 42] that:

$$P_{11'A'} := P_{BAB'A'}\iota^B\iota^A\iota^{B'} = 0. \tag{5.5}$$

Then, from (5.1) we obtain

$$\nabla_{11'}\pi_{A'} = -\mathrm{i}P_{11'A'}\xi = 0 \tag{5.6}$$

and, in particular,

$$\mathrm{d}\pi_{1'}/\mathrm{d}u = 0. \tag{5.7}$$

From this and (5.4) we obtain the formula,

$$u(\mathrm{P}) = u(\mathrm{Q}) - \mathrm{i}\xi/\pi_{1'} \tag{5.8}$$

for the location of P in terms of that of Q and the value $(\xi\iota^A, \pi_{A'})$ at Q. If (5.3) is *not* satisfied, then this formula would give us a *complex* value of u on γ. Thus we may think of the non-null twistors as intersecting I^+ in points with complex u coordinates. That is, they intersect the "complexified" I^+. This point of view will be developed further shortly.

Now consider what happens when $\pi_{A'}$ is proportional[31] to $\iota_{A'}$, i.e. when $\pi_{1'} = 0$. We have, at each point of γ, a local twistor of the form

$$(\xi\iota^A, \eta\iota_{A'}), \tag{5.9}$$

where ξ and η are constant all along γ. It is worth examining the meaning of such a quantity in the case of Minkowski space. Here I^+ may be viewed as a light cone with a vertex, namely I. A local null twistor $(\xi\iota^A, \pi_{A'})$ at Q, which approaches the form $(\xi\iota^A, \eta\iota_{A'})$ at Q in the limit, represents a global twistor Z^α where $\mathrm{P} = \mathrm{Z} \cap \gamma$ approaches I in the limit (see (5.8): the value $u(\mathrm{P})$ approaches $\pm\infty$). Thus, $(\xi\iota^A, \eta\iota_{A'})$ at Q represents some null twistor Z^α whose null line passes through I. But we could represent this same null twistor Z^α equally well using any *other* generator of I^+. In particular, since Z itself now becomes a generator, we may use local twistors along Z to describe Z^α, these having the particular form

$$(0, \eta\iota_{A'}). \tag{5.10}$$

Our description of Minkowski space twistors in terms of I^+ thus entails a further identification to be made whenever the local twistors we use take the form (5.9). Such an identification *can*, in fact, be carried out for asymptotically flat space-times (although it may not always be desirable to do so). How then do we carry out such an identification in general? There are various equivalent ways of achieving this, but they all amount to the following. Specialize the choice of a conformal factor further so that the generators of I^+ are not only parallel but so that in addition, the cross-sections of I^+ are all spheres. (This can be done [41].) Then use a form of *modified* local twistor transport in which the parts of $P_{ABA'B'}$ of the form

$$P_{11A'B'} \qquad \text{and} \qquad P_{AB1'1'} \tag{5.11}$$

are formally put equal to zero while those of the form

$$P_{1A1'A'}, \qquad P_{A11'A'}, \qquad P_{1AA'1'}, \qquad P_{A1A'1'} \tag{5.12}$$

are retained. (There is no inconsistency between these requirements because of (5.5).) When applied to local twistors of the form (5.9) this transport turns out to be integrable, so the required equivalence relation between quantities (5.9) at the various points of I^+ may thereby be achieved. The resulting equivalence classes are called *asymptotic twistors entirely on* I^+. The equivalence classes defined earlier for local twistors on I^+ of the form (5.2) which are *not* of the form (5.9), are the *(future) asymptotic twistors not entirely on* I^+. The *(future) asymptotic twistors* are simply the quantities of one or other of these kinds.

---

[31] There is no invariant significance in the further specialization to $\pi_{A'} = 0$ since, assuming $\xi \neq 0$, this condition is not invariant under conformal rescaling, cf. (3.1) even when the rescaling factor is chosen constant along generators of I^+.

The above procedure may seem to the reader to be artificial. However, this is not really so. We may think of it in another way. We may assume that the gravitational radiation emitted (or transmitted) is of an effectively finite duration, that is to say that it falls off appropriately to zero when the absolute value of u is large. Now the quantities (5.11) describe, in effect, the Bondi–Sachs news function N (and $\bar{N}$) and so must be expected to approach zero for large $|u|$. The quantities (5.12) are, on the other hand, constant along the generators of $\mathscr{I}^+$ (and are, in fact, constant multiples of $\iota_A \iota_{A'}$, over the whole of $\mathscr{I}^+$). In effect, we may think of the modified local twistor transport as being the limit as $|u| \to \infty$ of ordinary local twistor transport on $\mathscr{I}^+$. In the limit, this local twistor transport becomes integrable for the quantities (5.9).

It is worth remarking, also, that there is a limiting case of our above construction in which things can be seen a little more simply. This is to make a choice of Ω in which the cross-sections of $\mathscr{I}^+$ are *planes* and to send one of the generators of $\mathscr{I}^+$ effectively "back to infinity". With this choice of Ω it is possible to scale ι^A so that it is constant over the whole of $\mathscr{I}^+$. In fact, we can arrange

$$\nabla_{AA'}\iota^B = 0. \tag{5.13}$$

We also have

$$P_{11A'B'} = N\iota_{A'}\iota_{B'} \tag{5.14}$$

with this choice of Ω, and the quantities (5.12) all vanish. Such a choice of Ω and ι^A is a very convenient one for calculations on $\mathscr{I}^+$ (cf. [42, 43]). Recall also that $\Psi_{ABCD} = 0$ on $\mathscr{I}^+$, so very few curvature components survive on $\mathscr{I}^+$, the complex quantity N being the most significant one. To construct our equivalence relation between local twistors of the form (5.9), we can proceed by making such a choice of Ω, so that at one point $Q \in \mathscr{I}^+$ a given local twistor is reduced to the form $(\xi\iota^A, 0)$. There is still just enough freedom in Ω for this, provided that $\xi \neq 0$. Then the other local twistors equivalent to this one are the ones also of this form $(\xi\iota^A, 0)$, where ξ is *constant* throughout $\mathscr{I}^+$. The null line represented by the resulting asymptotic twistor is then the generator of $\mathscr{I}^+$ which is sent back to infinity with this choice of Ω.

In the case of Minkowski space, the asymptotic twistors which are entirely on $\mathscr{I}^+$ are those (apart from the zero twistor) whose representations in the C-picture are as points of the line I. If we do not carry out the identification, over the *whole* of $\mathscr{I}^+$, between local twistors of the form (5.9), but use only the previous identification given by local twistor transport along generators (which applied also to (5.2)), then we obtain a *blown-up asymptotic twistor* entirely on $\mathscr{I}^+$. This terminology comes from the name of a procedure sometimes applied in algebraic geometry in which a certain set of points may, in an algebraic fashion, be replaced by another set of higher dimensionality. In this case, I is "blown-up" to become a quadric surface, each point of which represents not only a point of I but also a plane through I.

Asymptotic twistor space may be viewed as a *fibre bundle* of, roughly speaking, spin-space over spin-space (actually over the dual complex conjugate spin-space). This generalizes the situation in Minkowski space-time, where, referred to an ordinary Minkowski metric and origin, a twistor Z^α may be represented as $(\omega^A, \pi_{A'})$. The base space is the space of $\pi_{A'}$ (namely the dual conjugate spin-space). The fibre over each fixed $\pi_{A'}$ is the space of ω^A (spin-space). This bundle structure is Poincaré (but not conformally) invariant. There is no Poincaré invariant cross-section, but the (Poincaré invariant) projection to the base is given by

$$(\omega^A, \pi_{A'}) \mapsto \pi_{A'}. \tag{5.15}$$

There is just one Poincaré invariant injection of spin-space into the bundle as a fibre, namely the fibre over $\pi_{A'} = 0$:

$$\omega^A \mapsto (\omega^A, 0). \tag{5.16}$$

This gives us a short exact sequence

$$0 \to S^A \to T^\alpha \to S_{A'} \to 0 \tag{5.17}$$

where the letters S and T represent spin-space and twistor-space, respectively, of the type indicated by the indices. Taking the dual spaces we get

$$0 \leftarrow S_A \leftarrow T_\alpha \leftarrow S^{A'} \leftarrow 0. \tag{5.18}$$

This exact sequence is actually also the complex conjugate of (5.17). All this structure persists[32] also for asymptotic twistors. To see this, we must identify the required spin-space – *asymptotic spin-space* – for an asymptotically flat space-time. This may be done in terms of a double covering of the cotangent bundle of the space of generators of $\mathscr{I}^+$. Equivalent, but more convenient for our present purposes, is to use the space of asymptotic twistors which are entirely on $\mathscr{I}^+$. Since we have an equivalence relation over the whole of $\mathscr{I}^+$, we can add two such asymptotic twistors simply by adding the corresponding local twistors at some given point Q of $\mathscr{I}^+$. The result is independent of Q. We can likewise multiply such asymptotic twistors by scalars and derive a two-complex-dimensional vector space structure. The two complex quantities ξ, η serve as components, where $(\xi\iota^A, \eta\iota_{A'})$ is the local twistor representation at Q. We can define the required (skew-symmetrical) spinor scalar product between $(\overset{1}{\xi}, \overset{1}{\eta})$ and $(\overset{2}{\xi}, \overset{2}{\eta})$ to be

$$\mathrm{i}\overset{1}{\xi}\overset{2}{\eta} - \mathrm{i}\overset{1}{\eta}\overset{2}{\xi} \tag{5.19}$$

provided ι^A is scaled so that

$$\nabla_{AA'}\Omega = -\iota_A \iota_{A'} \tag{5.20}$$

on $\mathscr{I}^+$. (Eq. (5.20) is in fact consistent with our other assumptions.) The defintion (5.19) turns out to be independent of Q and of the choice of Ω (because the scaling (5.20) implies that $\iota^A \iota_{A'}$ is independent of Ω). This invariant skew scalar product completes the definition of the structure of asymptotic spin-space. The definition does in fact agree with the ordinary spin-space structure for Minkowski space.

Since we have *defined* asymptotic spin-space as a subspace of asymptotic twistor space, the injection (5.16) is established (as an inclusion) for asymptotically flat space-time. Now examine the projection (5.15). This can be re-expressed

$$(\omega^A, \pi_{A'}) \mapsto (0, \pi^{A'}) \tag{5.21}$$

where the expression on the right refers to a twistor of valence $\left[{0 \atop 1}\right]$ and is obtained by

---

[32] Except that the interpretation of the complex conjugate of asymptotic twistor space as being also its dual space may not be valid.

$$Z^\alpha \mapsto I_{\alpha\beta} Z^\beta. \tag{5.22}$$

Reverting, now, to the description of asymptotic twistors in terms of local twistors on $\mathscr{I}^+$, we can derive the form (at *any* Q)

$$I_{\alpha\beta} \leftrightarrow \begin{pmatrix} 0 & -\mathrm{i}\iota_A \iota^{B'} \\ \mathrm{i}\iota^{A'}\iota_B & 0 \end{pmatrix} \tag{5.23}$$

for the infinity twistor. Applying (5.23) to (5.22), we obtain, with respect to Q, the version of (5.21) which applies to asymptotic twistors:

$$(\xi\iota^A, \pi_{A'}) \mapsto (-\mathrm{i}\pi_{1'}\iota_A, 0) \tag{5.24}$$

(Having the same form as the complex conjugate of (5.10), the right-hand side of (5.24) describes an asymptotic twistor of valence $[{}^0_1]$ as required.) The fibre above the asymptotic spinor given by the right-hand side of (5.24) is the kernel of (5.24), that is, the space of $(\xi\iota^A, \pi_{A'})$ for $\pi_{1'} = 0$. We can use ξ and $\pi_{0'}$, at Q, as components for this two-complex-dimensional vector space. There is no invariant skew-symmetrical scalar product. Instead, the group of the fibre may be taken to be simply $\mathrm{L}(2, \mathbb{C})$.

We can add two asymptotic twistors if they lie in the same fibre, but there is no clear way to do so if they are associated with different generators of $\mathscr{I}^+$. Any asymptotic twistor can be multiplied by a scalar.

Let us now consider how the complex structure for the space of asymptotic twistors (or of blown-up asymptotic twistors) is to be defined. The clearest way to do this seems to be to give an interpretation of asymptotic twistors in terms of the *complexified* $\mathscr{I}^+$. This is not the most satisfactory procedure mathematically, since in order that this complexification be possible at all, it is necessary that the structure of $\mathscr{I}^+$ be *analytic*. This entails that the outgoing gravitational radiation must be analytic, so gravitational pulse waves or sandwich waves (e.g. Bondi's "time-bomb") must be excluded. It would be more satisfactory to define all the required structure in terms of the *real* $\mathscr{I}^+$ – and presumably, it will eventually become possible to do this – but in any case the loss of generality involved in our approach may be small, since any non-analytic behaviour can be approximated arbitrarily closely by means of an analytic $\mathscr{I}^+$.

Let us use coordinates u, ζ for $\mathscr{I}^+$, where u is a real parameter define along the generators of $\mathscr{I}^+$ as described earlier (so u is actually a retarded time parameter which is constant along each generator and which labels the generators in the standard way, i.e. $\zeta = \exp(-\mathrm{i}\varphi)\cot\frac{1}{2}\theta$, where θ and φ are ordinary spherical polar coordinates for the spheres of cross-section of $\mathscr{I}^+$). To complexify $\mathscr{I}^+$, we allow u to become a complex parameter v and we "free" ζ from its complex conjugate $\bar{\zeta}$. That is to say, given any quantity, defined on $\mathscr{I}^+$ as an analytic function $\chi(u, \zeta, \bar{\zeta})$, we extend this to the complexified $\mathscr{I}^+$ – denoted $\mathbb{C}\mathscr{I}^+$ – by the unique (locally at least) holomorphic function of three independent complex variables v, ζ, $\tilde{\zeta}$, (still written $\chi(v, \zeta, \tilde{\zeta})$) whose value is given by $\chi(u, \zeta, \bar{\zeta})$ when $v = u$ is real and $\tilde{\zeta} = \bar{\zeta}$.

Since the range of ζ is really to be thought of as the Riemann sphere S^2 ($\zeta = \infty$ being allowed), this gives the topology of $\mathbb{C}\mathscr{I}^+$ as $\mathbb{R}^2 \times S^2 \times S^2$. (The parameter v ranges over $\mathbb{C} \cong \mathbb{R}^2$ and ζ and $\tilde{\zeta}$ each independently range over S^2.) This picture is not really quite accurate, however. This is because certain quantities defining the structure of $\mathscr{I}^+$ will become *singular* if defined too far away from the real section $\mathscr{I}^+$ of $\mathbb{C}\mathscr{I}^+$. This will apply, in particular, to the gravitational radiation field

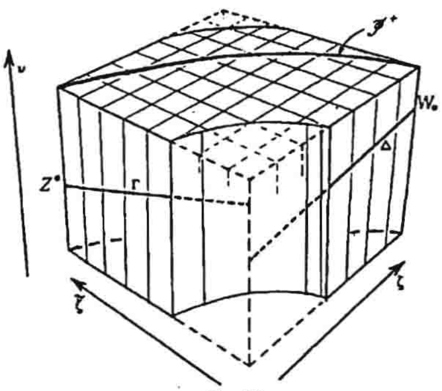


Fig. 40.

Ψ_4. (Functions, such as $(1+u^2)^{-1}$, which are well-behaved for real u become singular when extended to complex values.) Thus, the manifold $\mathbb{C}\mathscr{I}^+$ consists just of some neighbourhood of $\mathscr{I}^+$ in $\mathbb{R}^2 \times S^2 \times S^2$ (see fig. 40).

Let us consider local twistors on $\mathbb{C}\mathscr{I}^+$ of the form $(\xi\iota^A, \pi_{A'})$ at $Q \in \mathscr{I}^+$. Since v is complex we can, provided $\pi_{1'} \neq 0$, use (5.8) to find a point P on the generator $\mathbb{C}\gamma$ of $\mathbb{C}\mathscr{I}^+$ (or of $\mathbb{R}^2 \times S^2 \times S^2$, if P is singular for $\mathbb{C}\mathscr{I}^+$) through Q, at which local twistor transport reduces ξ to zero to give

$$(0, \pi_{A'}). \tag{5.25}$$

Now, starting from any point on $\mathbb{C}\mathscr{I}^+$, or on $\mathbb{C}M$, at which we have a local twistor of the form (5.25) we can, by means of local twistor transport, maintain the form (5.25) by moving in any (complex) direction of the type

$$t^a = \tau^A \pi^{A'} \tag{5.26}$$

(by (5.1)). In flat (or conformally flat) space, local twistor transport is integrable and defines a global twistor Z^α (section 3.1). The points at which the form (5.25) holds then constitute a complexified plane (a four-real-dimensional subset of $\mathbb{C}M$) whose tangent vectors are all of the type (5.26). In the case of a null twistor Z^α, such a plane contains one real null line of points (namely the line Z); in the general case, when $Z^\alpha \bar{Z}_\alpha \neq 0$, no point on the plane is real. This complexified plane (assuming that it is not contained in $\mathbb{C}\mathscr{I}^+$) will intersect $\mathbb{C}\mathscr{I}^+$ in a complexified line $\Gamma = \mathbb{C}Z \cap \mathbb{C}\mathscr{I}^+$, the tangents to Γ being proportional to

$$t^a = \iota^A \pi^{A'} \tag{5.27}$$

since these are the complex vectors of the type (5.26) which are tangent to $\mathbb{C}\mathscr{I}^+$. In fact Γ is a *null geodesic* on $\mathbb{C}\mathscr{I}^+$, which follows from the fact that

$$\iota^B \pi^{B'} \nabla_{BB'} \pi_{A'} = 0 \tag{5.28}$$

(by (5.1)) together with (5.13). (If we do not choose to scale things so that (5.13) holds, then we obtain that $t^b \nabla_b t^a$ is proportional to t^a.) Recall that the concept of a null geodesic is conformally invariant—this applies also to complex null geodesics.

Reverting to the general case, when $\mathscr{M}$ is asymptotically flat, we can likewise define the null geodesic Γ through $P \in \mathbb{C}\mathscr{I}^+$ in the direction given by $\iota^A\pi^{A'}$, and associate the local twistor $(0, \pi_{A'})$ with each point of Γ, $(0, \pi_{A'})$ being carried along Γ by local twistor transport. Now the surfaces ζ = const. on $\mathbb{C}\mathscr{I}^+$ have tangent vectors of the type $\iota^A\kappa^{A'}$, as in (5.27); the surfaces $\tilde{\zeta}$ = const. have tangent vectors of the type $\tau^A\iota^{A'}$. Thus, Γ lies entirely in one ζ = const. surface. Each point of Γ is labelled by a different value of $\tilde{\zeta}$ (assuming $\pi_{A'}$ is not proportional to $\iota_{A'}$). The value of u at $\tilde{\zeta} = \bar{\zeta}$, together with $\pi_{A'}$ at that point serves to define Γ. Conversely, Γ, together with the scaling afforded by the choice of tangent vector $\iota^A\pi^{A'}$, defines u, ζ and $\pi_{A'}$, and so corresponds uniquely to an asymptotic twistor not entirely on $\mathscr{I}^+$. There are also null geodesics on $\mathbb{C}\mathscr{I}^+$ lying on $\tilde{\zeta}$ = const. planes. These have tangent vectors of the type $\mu^A\iota^{A'}$ and may, in a similar way, be associated with dual asymptotic twistors not entirely on $\mathscr{I}^+$. All null geodesics on $\mathbb{C}\mathscr{I}^+$ have one or other of these forms – except that there are some, namely the generators of $\mathbb{C}\mathscr{I}^+$, which have both forms simultaneously. These generators are associated with blown-up asymptotic twistors (entirely on $\mathscr{I}^+$). The asymptotic twistors entirely on $\mathscr{I}^+$ which are not blown up must be associated with whole $\tilde{\zeta}$ = constant planes; the dual asymptotic twistors of the same type are associated with ζ = const. planes.

This description of asymptotic twistors is given on $\mathbb{C}\mathscr{I}^+$ completely without reference to the notion of complex conjugate. It follows that the manifold of these asymptotic twistors has a naturally defined complex structure. This is the complex structure that we assign to asymptotic twistor space and which makes this space into a complex manifold. There is a snag, however. This is that for some asymptotic twistors, the point P that we defined to start with to obtain the form (5.25) may not actually lie in $\mathbb{C}\mathscr{I}^+$, but in the region excluded from $\mathbb{R}^2 \times S^2 \times S^2$ because $\mathbb{C}\mathscr{I}^+$ becomes singular there. It seems that in general the complex structure for asymptotic twistor space must break down in such regions. However, this only occurs outside some neighbourhood of the space of null asymptotic twistors. But it does have the effect that although the space of asymptotic twistors is compact, the subspace on which a complex structure is defined is non-compact. (It seems rather probable that this non-compact complex manifold can be extended to become compact again, but this is not certain. If this can be done, then it can be done uniquely and a well-defined symmetry group isomorphic to the complex Poincaré group or complex conformal group will arise. This is presently under investigation.)

To obtain the geometric meaning of the complex structure of asymptotic twistor space, let us consider the asymptotic analogue of the Kerr theorem[33]. The asymptotic C-picture consists of equivalence classes of proportional asymptotic twistors. Each point of C – I represents a null geodesic on $\mathbb{C}\mathscr{I}^+$ (in a ζ = const. plane but which is not a generator). A complex analytic surface in C defines a complex analytic congruence of such null geodesics. Those which meet $\mathscr{I}^+$ (represented by points on N) define a field on $\pi_{A'}$'s (up to proportionality) on $\mathscr{I}^+$. Now the equation (5.28) must be satisfied – or, if we are not concerned with the scaling of $\pi_{A'}$, but merely its flagpole direction – the equation

$$\iota^B\pi^{B'}\pi^{A'}\nabla_{BB'}\pi_{A'} = 0. \tag{5.29}$$

---

[33] We are deeply indebted to E.T. Newman for informing us of his work with Aronson, Lind and Messmer [44] on asymptotically shear-free congruences. It was their asymptotic version of the Kerr theorem which provided the initial inspiration for the theory of asymptotic twistors described here. Their ϕ-functions correspond to complex-analytic functions on $\mathbb{C}\mathscr{I}^+$ whose level surfaces are ruled by null geodesics in ζ = const.

If we define $\pi_{A'}$ away from $\mathcal{I}^+$ by means of parallel propagation along the null geodesics (λ) in $\mathcal{M}$ which intersect $\mathcal{I}^+$ in the directions defined by the $\pi_{A'}$'s (i.e. $\bar{\pi}^B \pi^{B'} \nabla_{BB'} \pi_{A'} = 0$) then we arrive at the equation

$$\pi^{B'} \pi^{A'} \nabla_{BB'} \pi_{A'} = 0$$

at $\mathcal{I}^+$. This will be recognized [9] as the condition that the congruence (λ) should be shear-free ($\sigma^0 = 0$) at $\mathcal{I}^+$ i.e. that it should be *asymptotically shear-free.*

Finally, let us define the scalar product on asymptotic twistor space. Let Z^α be an asymptotic twistor and W_β a dual asymptotic twistor, neither of which lies entirely on $\mathcal{I}^+$. Let Z^α be associated with the null geodesic Γ on $\zeta = \zeta_1$ on $\mathbb{C}\mathcal{I}^+$ and W_β associated with the null geodesic Δ on $\tilde{\zeta} = \tilde{\zeta}_2$ on $\mathbb{C}\mathcal{I}^+$. There will be a unique generator of $\mathbb{C}\mathcal{I}^+$ which meets both of Γ and Δ, namely that given by $\zeta = \zeta_1, \tilde{\zeta} = \tilde{\zeta}_2$, provided that the intersection points do not lie in the singular region. If they do (see fig. 40), the scalar product has become singular, so suppose they do not. Then (providing also that the singular region of $\mathbb{C}\mathcal{I}^+$ does not disconnect the two points) we can use local twistor transport along this generator to carry the local twistor defining W_β to the intersection with Γ and perform the scalar product at that point (or else we can carry the local twistor defining Z^α to the intersection with Δ – it makes no difference to the answer).

When $W_\beta = \bar{Z}_\beta$, this gives us the definition of $Z^\alpha \bar{Z}_\alpha$ which comes directly from the local twistor definition of $Z^\alpha \bar{Z}_\alpha$ implicit in (5.3). (In fact, the above scalar product of Z^α with W_β is simply the "complexification" of $Z^\alpha \bar{Z}_\alpha$ where $\bar{Z}_\alpha$ is allowed to become "freed" from Z^α.) Having a real-valued scalar function defined on asymptotic twistor space as well as a complex structure, we have also a Kähler manifold structure (of signature (++−−) [45]. This leads to a concept of covariant derivative and of curvature for asymptotic twistor space. This Kähler curvature can be computed explicitly in terms of quantities relating to the asymptotic space-time curvature. These quantities are Ψ_4^0, Ψ_3^0, $\mathrm{Im}\,(\Psi_2^0)$ cf. [9], these being the components of

$$\mathrm{i}\Psi_{AB11}\iota_{A'}\iota_{B'} - \mathrm{i}\,\iota_A \iota_B \bar{\Psi}_{A'B'1'1'},$$

and (in the case of non-null twistors) certain integrals of such quantities up the generators of $\mathcal{I}^+$. These calculations are too involved to be usefully described here.

5.2. *Scattering off a fixed source*

In section 3.2 we considered the classical scattering of zero rest-mass particles (described in terms of null twistors) by a fixed gravitational or electromagnetic wave. Employing the Hamiltonian equations which arose we attempted to find the correct twistor description of the corresponding quantum processes. In the electromagnetic case this was successful only after the (unmotivated) introduction of an extra line into the diagrams; in the gravitational case success has not yet been achieved. In this section we briefly consider a variation on this programme in which the scattering is effected by the field of a fixed source rather than by a free wave. The contour integrals which arise present some distinctly new features – and are in several respects more satisfactory than the ones arising in the case of the free waves. The results obtained for the classical scattering turn out to be correct, but a valid procedure for obtaining the correct quantum scatterings at first proved to be elusive. This narrowed down the area in which new developments in the theory were required. The twistor description of the electromagnetic and gravitational potentials given in section 5.3 in fact forms the basis of these new developments.

In order to describe the electromagnetic or (linearized) gravitational field of a massive (spinning) source in twistor terms, it will be convenient first to discuss the 2-index twistor $A^{\alpha\beta}$ which represents its energy-momentum and angular momentum. Generally, let P_a be the momentum and M^{ab} the angular momentum with respect to an origin O, of some finite physical system in special relativity. Normally P_a will be timelike (future-pointing) so the twistor description of sections 1.3 and 2.1 does not directly apply. But we can always represent (P_a, M^{ab}) for such cases (in many ways) as a *sum* of a finite number of (P_a, M^{ab}) of the type for which the twistor description (1.32), (1.34) *does* directly apply – i.e. for which the momentum is null and future-pointing and with the spin-vector proportional to it. In fact, we need only *two*, $(\overset{1}{P}_a, \overset{1}{M}{}^{ab})$ and $(\overset{2}{P}_a, \overset{2}{M}{}^{ab})$, say, so

$$P_a = \overset{1}{P}_a + \overset{2}{P}_a, \qquad M^{ab} = \overset{1}{M}{}^{ab} + \overset{2}{M}{}^{ab} \tag{5.30}$$

where $(\overset{i}{P}_a, \overset{i}{M}{}^{ab})$ is constructed from twistors $(\overset{i}{\omega}{}^A, \overset{i}{\pi}_{A'})$ according to (1.32) and (1.34). This can be achieved in many ways as is not hard to verify explicitly.

Now, the expressions (1.32) and (1.34) for the momentum and angular momentum in terms of a single twistor $Z^\alpha \leftrightarrow (\omega^A, \pi_{A'})$ are

$$P_a = \bar{\pi}_A \pi_{A'}, \qquad M^{ab} = \mathrm{i}\omega^{(A}\bar{\pi}^{B)}\epsilon^{A'B'} - \mathrm{i}\epsilon^{AB}\bar{\omega}^{(A'}\pi^{B')}. \tag{5.31}$$

These quantities are equivalent to a *subset* of the components of $Z^\alpha \bar{Z}_\beta$. They are defined once $\pi_{A'}\bar{\pi}^B$ and $\omega^{(A}\bar{\pi}^{B)}$ are known, i.e. once the symmetrized outer product of $(\omega^A, \pi_{A'})$ with $(\bar{\pi}^B, 0)$ is known. We have $I^{\beta\gamma}\bar{Z}_\gamma \leftrightarrow (\bar{\pi}^B, 0)$ (compare (5.22)) so the *angular momentum twistor*

$$A^{\alpha\beta} = 2Z^{(\alpha}I^{\beta)\gamma}\bar{Z}_\gamma \leftrightarrow \begin{pmatrix} 2\omega^{(A}\bar{\pi}^{B)} & \bar{\pi}^A \pi_{B'} \\ \pi_{A'}\bar{\pi}^B & 0 \end{pmatrix} \tag{5.32}$$

may be used to describe the 4-momentum and angular-momentum in the null case. In the general timelike case we can extend this by linearity. Thus, the angular momentum twistor, quite generally, is

$$A^{\alpha\beta} \leftrightarrow \begin{pmatrix} -2\mathrm{i}\mu^{AB} & P^A{}_{B'} \\ P_{A'}{}^B & 0 \end{pmatrix} \tag{5.33}$$

where

$$M^{ab} = \mu^{AB}\epsilon^{A'B'} + \epsilon^{AB}\mu^{A'B'}; \qquad \mu^{AB} = \mu^{BA}. \tag{5.34}$$

We have $A^{\alpha\beta}$ of the form

$$A^{\alpha\beta} = 2E^{(\alpha}_\gamma I^{\beta)\gamma} \tag{5.35}$$

where

$$E^\alpha_\beta = \sum_i \overset{i}{Z}{}^\alpha \overset{i}{\bar{Z}}_\beta \tag{5.36}$$

the various $\overset{i}{Z}{}^\alpha$ being the twistors representing the different null-momentum systems into which our general system has been decomposed (e.g. for (5.30) we have $E^\alpha_\beta = \overset{1}{Z}{}^\alpha \overset{1}{\bar{Z}}_\beta + \overset{2}{Z}{}^\alpha \overset{2}{\bar{Z}}_\beta$). Note that E^α_β has the Hermitian property

$$\overline{E}{}^{\alpha}_{\beta} = E^{\alpha}_{\beta}. \tag{5.37}$$

(The left-hand side is, of course, $\overline{E^{\beta}_{\alpha}}$.) It also has the semi-definiteness property

$$W_{\alpha} E^{\alpha}_{\beta} \overline{W}{}^{\beta} \geqslant 0 \qquad \text{for all } W_{\alpha} \tag{5.38}$$

as follows at once from (5.36). The form (5.35), together with (5.37), implies

$$A^{\alpha\beta} = A^{\beta\alpha}, \qquad A^{\alpha\gamma} I_{\beta\gamma} = \overline{A}_{\beta\gamma} I^{\alpha\gamma}. \tag{5.39}$$

Conversely, (5.39) implies the existence of E^{α}_{β} for which (5.35) and (5.37) hold. The semi-definiteness (5.38) implies that

$$V_{\alpha} A^{\alpha\gamma} I_{\beta\gamma} \overline{V}{}^{\beta} \geqslant 0 \qquad \text{for all } V_{\alpha}. \tag{5.40}$$

Indeed, the second relation in (5.39) is also implied by (5.40) and the *existence* of E^{α}_{β}, subject to (5.38), is also implied by this.

Note that removal of the trace from E^{α}_{β}:

$$E^{\alpha}_{\beta} \to E^{\alpha}_{\beta} - \tfrac{1}{4} E^{\gamma}_{\gamma} \delta^{\alpha}_{\beta} \tag{5.41}$$

makes no difference to (5.35) (owing to the skew-symmetry of $I^{\alpha\beta}$) although the relation (5.38) would be affected. Given $A^{\alpha\beta}$ satisfying (5.39), the choice of E^{α}_{β} in (5.35) is clearly very far from unique. In fact, the "gauge freedom" in E^{α}_{β} is

$$E^{\alpha}_{\beta} \to E^{\alpha}_{\beta} + S^{\alpha\gamma} I_{\beta\gamma} + \overline{S}_{\beta\gamma} I^{\alpha\gamma} \tag{5.42}$$

(ignoring the semi-definiteness (5.38)) where

$$S^{\alpha\beta} = -S^{\beta\alpha}. \tag{5.43}$$

Let us now consider the quadratic function

$$A^{\alpha\beta} W_{\alpha} W_{\beta} \tag{5.44}$$

defined in dual twistor space. Since this is holomorphic and homogeneous, its vanishing defines, by Kerr's theorem (cf. [5]), the shear-free congruence of null straight lines

$$\{W : A^{\alpha\beta} W_{\alpha} W_{\beta} = 0 = W_{\alpha} \overline{W}{}^{\alpha},\ W_{\alpha} \neq 0\}. \tag{5.45}$$

When the spin vector (1.29) vanishes, this congruence consists of all the null lines which intersect the world line of the relativistic mass centre (see section 1.3). In the general case of a system with spin, the congruence is of the more general kind considered by Kerr [47] (in connection with his vacuum solution of Einstein's equations representing a rotating black hole). The congruence possesses a local rotation if and only if the system has a non-zero spin. It also turns out that the null directions, at each point in space-time, which are tangent to the lines of the congruence are precisely the "principal null directions" [9, 46] of the angular momentum tensor M^{ab}. (These are the flagpole directions of λ^{A} and ν^{A}, where $\mu^{AB} = \lambda^{(A} \nu^{B)}$.)

It is the function (5.44) which also enters into the definition of a passive holomorphic function generating a solution of Maxwell's equations surrounding a charged spinning (i.e. with magnetic moment) source, or a solution of the linearized Einstein equations (the linearized Schwarzschild or Kerr solution) for a massive spinning source. In fact, putting

$$f(W) = (A^{\alpha\beta}W_\alpha W_\beta)^{-2} \qquad \text{and } g(W) = (A^{\alpha\beta}W_\alpha W_\beta)^{-3}$$

we have f homogeneous of degree -4 (as required for the construction of a Maxwell field) and g homogeneous of degree -6 (as required for the construction of a linearized gravitational field). We have already noted, in section 2.3, that when $g(W)$ is the inverse cube of a quadratic form, then the resulting linearized gravitational field has algebraic type $\{2,2\}$ – as happens to be the case for a (linearized) Schwarzschild or Kerr solution. It is indeed true that the functions do yield fields of stationary sources of the required type. When the rotation vanishes, we get monopole solutions (Coulomb in the electromagnetic case and linearized Schwarzschild in the gravitational case).

Let $B_{\alpha\beta}$ be inverse to $A^{\alpha\beta}$,

$$A^{\alpha\beta}B_{\beta\gamma} = \delta^\alpha_\gamma, \tag{5.46}$$

and put

$$\Delta = \{\det A^{\alpha\beta}\}^{1/2} = \{\det B_{\alpha\beta}\}^{-1/2}. \tag{5.47}$$

(Only in the case of a null momentum is $A^{\alpha\beta}$ singular. We shall ignore this case henceforth.) One verifies at once that the properties

$$B_{\alpha\beta} = B_{\beta\alpha}, \qquad B_{\alpha\gamma}I^{\beta\gamma} = \bar{B}^{\beta\gamma}I_{\alpha\gamma} \tag{5.48}$$

and

$$X^\alpha B_{\alpha\gamma}I^{\beta\gamma}\bar{X}_\beta \geqslant 0 \qquad \text{for all } X^\alpha \tag{5.49}$$

follow directly from (5.46) and the corresponding properties (5.39) and (5.40) for $A^{\alpha\beta}$. Also Δ is *real*:

$$\Delta = \bar{\Delta}. \tag{5.50}$$

The condition for the absence of rotation turns out to be

$$\bar{A}_{\alpha\beta} = B_{\alpha\beta}\Delta. \tag{5.51}$$

The charge integral for an electromagnetic field was briefly mentioned in section 3.3. This was $\oint f(W)\mathcal{D}W$ (with some constant multiplying factor). Here we have

$$\epsilon = \oint f(W)\mathcal{D}W = \pi^2\Delta^{-1}. \tag{5.52}$$

The contour can be conveniently chosen to be an S^3. Similarly, the integral giving the energy-momentum and angular momentum for the source of a linearized gravitational field is $\oint W_\alpha W_\beta g(W)\mathcal{D}W$ which here gives:

$$\oint W_\alpha W_\beta g(W)\mathcal{D}W = \tfrac{1}{4}\pi^2 B_{\alpha\beta}\Delta^{-1} \tag{5.53}$$

(although here Δ^{-1} measures the rest-mass rather than the charge). This may be obtained by differentiating (5.52) with respect to $A^{\alpha\beta}$ (temporarily dropping the constraints (5.39)) since

$$\partial f/\partial A^{\alpha\beta} = -2W_\alpha W_\beta g \tag{5.54}$$

and

$$\frac{\partial\Delta}{\partial A^{\alpha\beta}} = \tfrac{1}{2}B_{\beta\alpha}\Delta, \qquad \frac{\partial B_{\alpha\beta}}{\partial A^{\gamma\delta}} = -B_{\alpha\gamma}B_{\delta\beta}. \tag{5.55}$$

Notice the reversal of roles of $A^{\alpha\beta}$ and $B_{\alpha\beta}$ here. The angular momentum twistor arising here is actually (proportional to) $\bar{B}^{\alpha\beta}$.

We wish to consider scattering of charged massless particles off the electromagnetic field whose passive twistor function is $f(W_\beta)$. To compute the Hamiltonian we need to evaluate the integral for the active function[34]:

$$\tilde{f}(Z^\alpha) = \oint (W\cdot Z)_0 f(W_\beta)\mathcal{D}W. \tag{5.56}$$

Since this involves the formal expression $(W\cdot Z)_0$ we had best differentiate with respect to Z^α:

$$\partial\tilde{f}/\partial Z^\alpha = \oint W_\alpha(W\cdot Z)_1 f(W)\mathcal{D}W. \tag{5.57}$$

This is now well-defined. The singularity structure is different from that of (5.52) since there is the extra pole arising from $(W\cdot Z)_1$. This pole gets in the way of the previous contour but, remarkably, a new contour emerges which links the new pole and is not homologous to zero. A similar phenomenon occurs again if another pole $(W\cdot Y)_1$ is inserted into the integral and once more if yet another pole $(W\cdot X)_1$ is inserted. (To ensure uniqueness of the contour up to homology class, we must insist that each pole is "linked" once in a "positive" sense.) It is worthwhile to list the results of some relevant integrals

$$\oint \frac{W_\alpha \mathcal{D}W}{W_\beta Z^\beta (A^{\rho\sigma}W_\rho W_\sigma)^2} = \frac{2\pi^2}{\Delta}\frac{B_{\alpha\beta}Z^\beta}{B_{\rho\sigma}Z^\rho Z^\sigma} \tag{5.59}$$

$$\oint \frac{\mathcal{D}W}{W_\alpha Y^\alpha W_\beta Z^\beta (A^{\rho\sigma}W_\rho W_\sigma)} = 8\pi^3\{2\epsilon_{\alpha\beta\gamma\delta}\epsilon_{\lambda\mu\rho\sigma}A^{\alpha\lambda}A^{\beta\mu}Y^\gamma Z^\delta Y^\rho Z^\sigma\}^{-1/2} \tag{5.60}$$

$$\oint \frac{W_\alpha \mathcal{D}W}{W_\beta X^\beta W_\gamma Y^\gamma W_\delta Z^\delta (A^{\rho\sigma}W_\rho W_\sigma)} = 8\pi^3 \mathrm{i}\frac{\epsilon_{\alpha\beta\gamma\delta}X^\beta Y^\gamma Z^\delta}{\epsilon_{\xi\lambda\mu\nu}\epsilon_{\eta\rho\sigma\tau}A^{\xi\eta}X^\lambda Y^\mu Z^\nu X^\rho Y^\sigma Z^\tau}\,. \tag{5.61}$$

The integral (5.59) supplies the answer to (5.57) which we can immediately integrate to obtain (formally)

$$\tilde{f}(Z) = \frac{\pi^2}{\Delta}(B_{\alpha\beta}Z^\alpha Z^\beta)_0 = \epsilon \log(B_{\alpha\beta}Z^\alpha Z^\beta). \tag{5.62}$$

Our twistor scattering Hamiltonian is

$$H(Z^\alpha, \bar{Z}_\beta) = \tilde{f}(Z^\alpha) + \bar{\tilde{f}}(\bar{Z}_\beta), \tag{5.63}$$

---

[34] It turns out that for functions of the singularity structure considered here (unlike the case of a free wave) this operation can be inverted. In fact, we have $\tilde{f} \propto f$.

which gives

$$\delta\bar{Z}_\alpha = 2\mathrm{i}\epsilon\, B_{\alpha\beta}Z^\beta / B_{\rho\sigma}Z^\rho Z^\sigma . \tag{5.64}$$

Inserting explicit coordinates into this formula we obtain the correct lowest order "$1/r$" deflection angle. Also (5.64) gives the correct lowest order time delay – which turns out to be a constant independent of the impact parameter.

In a similar way the gravitational case can be treated. The final result is

$$\delta\bar{Z}_\alpha \propto \frac{1}{\Delta}\left\{B_{\alpha\gamma}I^{\beta\gamma}\bar{Z}_\beta \log|B_{\rho\sigma}Z^\rho Z^\sigma|^2 + \frac{B_{\alpha\beta}Z^\beta B_{\rho\sigma}\Gamma^\sigma Z^\rho \bar{Z}_\tau}{B_{\lambda\mu}Z^\lambda Z^\mu}\right\}$$

which must be considered formal, to some extent, because of the logarithm term. The logarithm only affects the time delay, the scattering angle being given by the second term. This second term is identical in form with that for the electromagnetic case except for the momentum factor

$$B_{\rho\sigma}\Gamma^\sigma Z^\rho \bar{Z}_\tau .$$

This simply ensures that the scattering angle is independent of the momentum of the scattered particle (since the whole term now has the same homogeneity degree as $\delta\bar{Z}_\alpha$) in accordance with the equivalence principle. Owing to the logarithm term there is no meaningful ("invariant") absolute concept of time delay. This is borne out by examination of the null orbits in the Schwarzschild solution. The time delay involves a similar logarithm term which diverges to minus infinity for large impact parameters.

Finally there is the question of how to obtain the correct quantum versions of these scattering problems using twistor methods. Blindly inserting the Hamiltonian into the formula $\langle p|q\rangle$ for the scalar product does not yield the correct matrix elements but gives zero instead. It appears that potentials must be brought in in order to obtain the correct results. A method of introducing potentials into twistor theory is given in the next section.

5.3. *Potentials and massive fields*

The fundamental formula (2.20) for a solution (in flat space) of the spin $\frac{1}{2}n$ zero-rest-mass field equation $\nabla^{PP'}\phi_{PR..S} = 0$, in terms of a general holomorphic function $f(W_\alpha)$ homogeneous of degree $-n-2$ in the components of the twistor W_α may be re-expressed as follows:

$$\phi_{PQ..S}(x^a) = \frac{1}{2\pi\mathrm{i}} \oint \lambda_P\lambda_Q ... \lambda_S f(\lambda_A, -\mathrm{i}x^{AA'}\lambda_A) D\lambda \tag{5.65}$$

where

$$D\lambda = \epsilon^{AB}\lambda_A \,\mathrm{d}\lambda_B$$

(cf. (3.35)). The verification that the zero rest-mass equations are indeed satisfied is now very direct since

$$\frac{\partial}{\partial x^{TT'}}\phi_{PQ..S} = \frac{1}{2\pi\mathrm{i}} \oint -\mathrm{i}\lambda_P\lambda_Q ... \lambda_S\lambda_T f_T(\lambda_A, -\mathrm{i}x^{AA'}\lambda_A) D\lambda$$

is manifestly symmetric in $PQ...ST$, where

$$\partial f/\partial W_\alpha \leftrightarrow (f^A, f_{A'}).$$

The formula (5.65) allows an immediate generalization to functions $F(W_\alpha; Z^\alpha)$, holomorphic in *two* twistors W_α, Z^α, and homogeneous of respective degrees $-p-2$ and $-q-2$. We can set

$$\alpha_{P...SG'..K'}(x^a) = \frac{1}{(2\pi \mathrm{i})^2} \oint \underbrace{\lambda_P...\lambda_S}_{p} \; \underbrace{\mu_{G'}...\mu_{K'}}_{q} \, F(\lambda_A, -\mathrm{i}x^{AA'}\lambda_A; \mathrm{i}x^{AA'}\mu_{A'}, \mu_{A'}) \mathcal{D}\lambda\mu \tag{5.66}$$

(where $\mathcal{D}\lambda\mu = \mathcal{D}\lambda \wedge \mathcal{D}\mu$ etc.; cf. (3.34)). Then, since $\partial\alpha.../\partial x^{TT'}$ is the sum of a term symmetric in its unprimed indices and a term symmetric in its primed indices, it follows that

$$\nabla^{PG'}\alpha_{PQ..SG'..K'} = 0. \tag{5.67}$$

Note that $\alpha_{...}$ depends on $F(W_\alpha; Z^\alpha)$ where $W_\alpha Z^\alpha = 0$ only. (One can also generalize in an obvious way to obtain two-point fields $\alpha_{P..SG'..K'}(x^a, y^b)$ satisfying $\partial\alpha_{P..SG'..K'}/\partial x^P_{P'} = 0$, $\partial\alpha_{P..SG'..K'}/\partial y^G_{G'} = 0$. These fields reduce to the previous ones if we set $y^a = x^a$, but they depend on $F(W_\alpha; Z^\alpha)$ also where $W_\alpha Z^\alpha \neq 0$.)

The equation (5.67) has considerable interest. When $p = q = 1$ it reduces to $\nabla^{PG'}\alpha_{PG'} = 0$, which is the Lorenz gauge condition if α_a is the electromagnetic potential. We can then define the electromagnetic field spinor by $\phi_{AB} = \nabla^{C'}_{A}\alpha_{BC'}$ and symmetry in AB is ensured. More generally, (5.67) implies that

$$\beta_{P..STH'..K'} = \nabla^{G'}_{T}\alpha_{P..SG'H'..K'} \tag{5.68}$$

is symmetric in $P...ST$ and that

$$\gamma_{Q..SG'..K'L'} = \nabla^{P}_{L'}\alpha_{PQ..SG'..K'} \tag{5.69}$$

is symmetric in $G'..K'L'$ where we use the fact that $\alpha_{P..SG'..K'}$ is symmetric in $P...S$ and in $G'...K'$. Furthermore, each of $\beta_{...}$ and $\gamma_{...}$ also satisfies (5.67) ($\nabla^{PH'}\beta_{P..TH'..K'} = 0$, $\nabla^{QG'}\gamma_{Q..SG'..L'} = 0$). Indeed, they each arise from a formula like (5.66), in which $F(W_\alpha; Z^\alpha)$ is replaced, respectively, by[35]

$$-\mathrm{i}I_{\rho\sigma}Z^\rho\,\partial F/\partial W_\sigma \qquad \text{or} \qquad \mathrm{i}I^{\rho\sigma}W_\rho\,\partial F/\partial Z^\sigma. \tag{5.70}$$

Note that the operation (5.68) increases the number of unprimed indices by one and decreases the number of primed indices by one, while the operation (5.69) does just the reverse. If we follow one operation by the other, the result is simply equivalent to application of $-\frac{1}{2}\Box$, where $\Box = \nabla^q\nabla_q$ is the flat space D'Alembertian. Thus, we have, for the twistor translation of the D'Alembertian:

$$\Box = -2\left(I_{\alpha\beta}Z^\alpha \frac{\partial}{\partial W_\beta}\right) \circ \left(I^{\rho\sigma}W_\rho \frac{\partial}{\partial Z^\sigma}\right) = -2I_{\alpha\beta}Z^\alpha I^{\rho\sigma}W_\rho \frac{\partial^2}{\partial W_\beta \partial Z^\sigma}. \tag{5.71}$$

---

[35] The observation that potential functions for zero rest-mass fields could be described by holomorphic functions of W_α and Z^α related by (5.70) had already been made by C.J.S. Clarke.

Employing the quantum mechanical replacements $-\partial/\partial Z^\alpha \leftrightarrow \bar{Z}_\alpha$, $\partial/\partial W_\beta \leftrightarrow \bar{W}^\beta$ and $\Box \leftrightarrow -m^2$, we obtain

$$m^2 = 2|I_{\alpha\beta}Z^\alpha \bar{W}^\beta|^2 . \tag{5.72}$$

It is reassuring that this agrees with the formula for the total mass of a system which, as in section 5.2, is composed of two systems each of zero rest-mass, namely that arising from Z^α and that arising from W_α. We have $P_a = \bar{Z}_A Z_{A'} + W_A \bar{W}_{A'}$, so $P_a P^a = 2|Z_{A'}\bar{W}^{A'}|^2$ whence (5.72) follows.

If $F(W_\alpha, Z^\alpha)$ is an eigenstate with eigenvalue $-m^2$ of the operation (5.71), then the resulting field $\alpha_{\dots}$ defined by (5.66) satisfies $(\Box + m^2)\,\alpha_{\dots} = 0$. This equation, together with (5.67), implies that $\alpha_{\dots}$ (paired up with $\beta_{\dots}$ of (5.68) or with $\gamma_{\dots}$ of (5.69) if desired) describes a Dirac irreducible "higher spin" free field of mass m (assumed positive, for the moment), and spin $\frac{1}{2}(p+q)$. (If $pq = 0$, then (5.67) becomes vacuous and is not needed. This applies, in particular to the normal Dirac equation for the electron, for which $p = 1, q = 0$ or $p = 0, q = 1$.) Thus, we have the potentiality to describe states of particles with non-zero rest-mass within the twistor formalism.

When $m = 0$ we have a way of describing the potentials directly in twistor terms. This gives a means of introducing interactions in a much more direct way than had been possible hitherto within the twistor formalism[36]. We are not now forced into considering only source-free fields, since the application of the operator (5.71) to F can give the source terms.

If $m = 0$ and sources are absent, it should be possible to obtain a function $F(W_\alpha)$ generating the field $\phi_{P..S}$ as in (5.65) from the function $F(W_\alpha; Z^\alpha)$ generating the same field as in (5.66) with $q = 0$. Provided $F(\lambda_A, -\mathrm{i}x^{AA'}\lambda_A; \mathrm{i}y^{AA'}\mu_{A'}, \mu_{A'})$ is independent of y^a, i.e. provided $I^{\alpha\beta}\partial F/\partial Z^\beta = 0$, we can write this

$$f(W_\alpha) = \frac{1}{2\pi\mathrm{i}}\oint F(W_\alpha; \mathrm{i}y^{AA'}\mu_{A'}, \mu_{A'})\mathcal{D}\mu.$$

References

[1] R. Penrose, in Combinatorial mathematics and its applications, ed. D.J.A. Welch (Academic Press, London, 1971).
[2] R. Penrose, Battelle Rencontres 1967, eds. C.M. de Witt and J.A. Wheeler (Benjamin, New York 1968).
[3] N.H. Kuiper, Ann Math. 50 (1949) 916;
H. Rudberg, Dissertation, University of Uppsala, Sweden 1958;
R. Penrose, in Proc. 1962 Conf. on Relativistic theories of gravitation, Warsaw (Polish Academy of Science, Warsaw 1965);
A. Uhlmann, Acta Phys. Polon. 24 (1963) 293.
[4] R. Penrose, Proc. Roy. Soc. (London) A 284 (1965) 159.
[5] R. Penrose, J. Math. Phys. 8 (1967) 345.
[6] E. Cartan, Ann. Ecole Norm. Super. 31 (1914) 263;
F. Klein, Gesammelte Math. Abh. (J. Springer, Berlin 1921);
H. Weyl, The classical groups (Princeton University Press, Princeton, New Jersey 1939);
R. Brauer and H. Weyl, Amer. J. Math. 57 (1935) 425.
[7] L.P. Eisenhart, Continuous groups of transformations (Princeton University Press 1933, reprinted by Dover 1961).
[8] H. Laue, Nuovo Cimento 38 (1971) 55.
[9] R. Penrose, Ann. Phys. (N.Y.) 10 (1960) 171;
E.T. Newman and R. Penrose, J. Math. Phys. 3 (1962) 566.

---

[36] In particular the Möller scattering formula (fig. 11) discussed in section 4.1 can now be completely justified.

[10] E.T. Whittaker, Proc. Roy. Soc. (London) A 158 (1937) 38;
W.T. Payne, Amer. J. Phys. 20 (1952) 253.
[11] R. Penrose, Null hypersurface initial data, in A.R.L. Techn. Report 63-56, ed. P.G. Bergmann (Office of Aerospace Research, U.S. Air Force 1963).
[12] P.A.M. Dirac, Proc. Roy. Soc. (London) A 155 (1936) 447;
M. Fierz and W. Pauli, Proc. Roy. Soc. (London) A 173 (1939) 211;
E.P. Wigner and V. Bargmann, Proc. Nat. Acad. Sci. (Washington) 34 (1948) 211.
[13] E. Cunningham, Proc. London Math. Soc. 8 (1910) 77;
H. Bateman, Proc. London Math. Soc. 8 (1910) 223;
J.A. McLennan Jr., Nuovo Cimento 10 (1956) 1360;
H.A. Buchdahl, Nuovo Cimento 11 (1959) 496.
[14] R.K. Sachs and P.G. Bergmann, Phys. Rev. 112 (1958) 674.
[15] B.L. van der Waerden, Nachrichten von der Gesellschaft der Wiss. Gottingen (1929) 100.
[16] W.A. Hepner, Nuovo Cimento 26 (1962) 351;
Y. Murai, Prog. Theoret. Phys. 9 (1953) 147; 11 (1954) 441 and Nucl. Phys. 6 (1958) 489;
I.E. Segal, Proc. Nat. Acad. Sci. (Washington) 57 (1967) 194.
[17] H. Bondi, Nature 186 (1960) 535;
H. Bondi, A.W.K. Metzner and M.J.G. van der Burg, Proc. Roy. Soc. (London) A 269 (1962) 21;
R.K. Sachs, Proc. Roy. Soc. (London) A 270 (1962) 103.
[18] E.T. Newman and T.W.J. Unti, J. Math. Phys. 3 (1962) 891.
[19] E. Grgin, Ph.D. thesis, Syracuse University (1966).
[20] J.A. Todd, Projective and analytic geometry (Pitman, London 1947);
J.G. Semple and L. Roth, Introduction to algebraic geometry (Clarendon Press, Oxford 1949).
[21] R. Penrose, Int. J. Theor. Phys. 1 (1968) 61.
[22] I. Robinson, J. Math. Phys. 2 (1961) 290.
[23] N. Kopczynski and L.S. Woronowicz, Reports Math. Phys. 2 (1971) 35.
[24] R. Penrose, J. Math. Phys. 10 (1969) 38.
[25] R.F. Streater and A.S. Wightman, *PCT*, spin statistics and all that (Benjamin, New York 1964).
[26] M. Crampin and F.A.E. Pirani, in Relativity and gravitation, eds. Ch.G. Kuper and A. Peres (Gordon and Breach, London 1971).
[27] N.J. Hicks, Notes on differential geometry (Van Nostrand, Princeton 1965).
[28] W.H. Brinkmann, Proc. Natl. Acad. Sci. (Washington) 9 (1923) 1; see also
J. Ehlers and W. Kundt, in Gravitation; an introduction to current research, ed. L. Witten (John Wiley, New York 1962) ch. 2.
[29] R. Penrose, General relativity (Papers in honour of J.L. Synge), ed. L. O'Raifertaigh for the Royal Irish Academy (Clarendon Press, Oxford 1972).
[30] M. Fierz, Helv. Phys. Acta 13 (1940) 45.
[31] I. Robinson and A. Trautman, Proc. Roy. Soc. (London) A 265 (1962) 463.
[32] A. Qadir, Ph.D. thesis, Birkbeck College, London (1971).
[33] R. Penrose, An analysis of the structure of spacetime (Adams Prize Essay, Princeton 1967).
[34] T. Fulton, F. Rohrlich and L. Witten, Rev. Mod. Phys. 34 (1962) 442.
[35] A. Trautman, F.A.E. Pirani and H. Bondi, Lectures on general relativity, Brandeis Summer Institute in Theoretical Physics, 1964, vol. 1 (Prentice-Hall, Englewood Cliffs, New Jersey 1965).
[36] L. Gross, J. Math. and Phys. 5 (1963) 687.
[37] R. Penrose, in Quantum theory and beyond, ed. E.T. Bastin (Cambridge University Press 1972).
[38] R. Penrose, in Magic without magic: John Archibald Wheeler, ed. J.R. Klauder (W.H. Freeman & Co., San Francisco 1972).
[39] M. Flato, J. Simon and D. Sternheimer, Ann. Phys. (N.Y.) 61 (1970) 78;
H. Kastrupp, Phys. Rev. 142 (1966) 1060.
[40] L. Infeld and B.L. van der Waerden, Sitz. ber. Preuss. Akad. 9 (1933) 380.
[41] R. Penrose, in Relativity, groups and topology, eds. C and B. deWitt, Les Houches lectures 1963 (Blackie and Son, London and Glasgow 1964).
[42] R. Penrose, in Relativity theory and astrophysics vol. 1, ed. J. Ehlers (American Mathematical Society, Providence, Rhode Island 1967).
[43] E.T. Newman and R. Penrose, Proc. Roy. Soc. (London) A 305 (1968) 175.
[44] B. Aronson, R. Lind, J. Messmer and E.T. Newman, J. Math. Phys. 12 (1971) 2462.
[45] S. Helgason, Differential geometry and symmetric spaces (Academic Press, New York 1962);
E. Nelson, Tensor analysis, Mathematical Notes (Princeton University Press 1967).
[46] J.L. Synge, Relativity; the special theory (North-Holland, Amsterdam 1956).
[47] R.P. Kerr, Phys. Rev. Letters 11 (1963) 237.
[48] J.D. Bjorken and S.D. Drell, Relativistic quantum fields (McGraw-Hill, New York 1965).

53

A) 'RADIATION, ENERGY AND THE BMS GROUP'; B) 'SOME RECENT DEVELOPMENTS'; C) 'N-P CONSTANTS'

3 summaries of lectures at International Centre for Theoretical Physics (Trieste) Topical meeting (1971) on Gravitation and Field Theory. ICTP, 1972

506 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants': 3 summaries of lectures at International Centre for Theoretical Physics (Trieste) Topical meeting (1971) on Gravitation and Field Theory

RADIATION, ENERGY AND THE BMS GROUP

R. Penrose

Birkbeck College, London, UK.

To calculate the mass loss due to gravitational radiation from an isolated system, it is necessary, owing to the non-locality of gravitational energy, to perform integrals over hypersurfaces extending all the way to infinity, or over surfaces lying at infinity. This is much facilitated by a device of "making infinity finite" with the introduction of a suitable conformal factor Ω. Accordingly, asymptotically flat space-time acquires a conformal boundary, defined by $\Omega = 0$, consisting of a smooth null hypersurface $\mathscr{I}^-$ representing past null infinity, a smooth null hypersurface $\mathscr{I}^+$ representing future null infinity and three points I^-, I^0 and I^+ (which, however, are singular in most cases) representing, respectively, past timelike, spacelike and future timelike infinities.

The utility of this picture rests on the conformal invariance of null geodesies and of zero rest-mass free fields. The gravitational radiation field is measured by a complex component Ψ_4^0 of the (scaled up) Weyl curvature field on $\mathscr{I}^+$ (outgoing field) or on $\mathscr{I}^-$ (incoming field). Bondi's news function N is a complex component of the Ricci tensor on $\mathscr{I}^\pm$ and turns out to be a time integral of the radiation field Ψ_4^0. The positive-definite quantity $N\bar{N}$ measures the flow of gravitational energy momentum across infinity, according to the Bondi-Sachs mass-loss formula. The total energy momentum at any one retarded time is measured by an integral of the quantity $\sigma^0 N - \Psi_2^0$ over a cross-section Σ of $\mathscr{I}^+$, where Ψ_2^0 is another (scaled) Weyl tensor component at $\mathscr{I}^+$ and where σ^0 is the asymptotic shear of the null (retarded time) hypersurface meeting $\mathscr{I}^+$ in Σ.

The Bondi-Metzner-Sachs group is an asymptotic symmetry group for asymptotically flat space-time. It may be expressed as the group of self-transformations of $\mathscr{I}^+$ preserving a certain ("strong") conformal intrinsic structure. The BMS group is an infinite parameter "Lie" group which contains the translations as a canonically defined normal subgroup. However, it also contains more general transformations known as supertranslations. The homogeneous Lorentz group appears canonically as a factor group of the BMS group by the supertranslation group. Many subgroups of the BMS group can be found which are isomorphic with the Poincaré group, but apparently no one of them can be singled out in a very natural way. Consequently no naturally defined concept of angular momentum for asymptotically flat space-times has yet emerged. The difficulty arises because the asymptotic shear structure σ^0 differs from that of Minkowski space when news function N is present.

REFERENCES

H. Bondi, M. van der Burg and A. Metzner, Proc. Roy. Soc. (London) A269, 21 (1962).

R.K. Sachs, Proc. Roy. Soc. (London) A270, 103 (1962).

R.K. Sachs, Phys. Rev. 128, 2851 (1962).

E.T. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962); 4, 998 (1963).

E.T. Newman and T. Unti, J. Math. Phys. 3, 891 (1962).

R. Penrose, Phys. Rev. Letters 10, 66 (1963).

R. Penrose, article in Relativity, Groups and Topology, eds. C.M. DeWitt and B.S. DeWitt (Gordon & Breach, New York 1964).

R. Penrose, Proc. Roy. Soc. (London) A284, 159((1965).

L. Tamburino and J. Winicour, Phys. Rev. 150, 1039 (1966).

V. Cantoni, J. Math. Phys. 8, 1700 (1967).

E.T. Newman and R. Penrose, J. Math. Phys. 7, 863 (1966).

R. Penrose, article in Battelle Recontres, eds. C.M. DeWitt and J.A. Wheeler (Benjamin Inc., New York 1968).

508 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants': 3 summaries of lectures at International Centre for Theoretical Physics (Trieste) Topical meeting (1971) on Gravitation and Field Theory

SOME RECENT DEVELOPMENTS

R. Penrose

Birkbeck College, London, UK.

Some recent work by Newman and his collaborators has led to an alternative approach to the question of asymptotic symmetry groups for asymptotically flat space-time. Although at first sight this new approach seems physically obscure, there are reasons for believing that it may ultimately have deeper significance than the BMS group approach. It involves choosing "canonical cross-sections" of $\mathscr{I}^+$ for which $\sigma^0 = 0$. However, this can only be achieved in general with complex cross-sections. The resulting asymptotic symmetry group which naturally arises appears to be the complex Poincaré group. This requires physical interpretation: for example, a complex displacement corresponds to the addition of an intrinsic angular momentum per unit mass.

These ideas find natural expression within the theory of twistors and allow that theory to be extended from flat to asymptotically flat space-times. The use of complex numbers and holomorphic functions plays a vital role in twistor theory. The geometrical manifestation of complex structure emerges in connection with Kerr's construction of the shear-free null congruences in flat space-time and in Newman's generalization of this to asymptotically flat space-time. The shear-free condition has importance for the question of clear-cut wave propagation and to the nature of scattering by gravitational waves. One of the main aims of twistor theory is to provide a new formalism for quantized gravitational theory. The formalism is not based on points but on null lines. Consequently, quantization should result in the "smearing" not of light cones but of space-time points.

Flat space twistors are the spinors for SO(2,4) which is locally isomorphic to the conformal group. A twistor Z^α (and its complex conjugate $\bar{Z}_\alpha$) describes the momentum and angular momentum structure of a zero rest-mass particle with intrinsic spin $\frac{1}{2}\hbar Z^\alpha \bar{Z}_\alpha$, but Z^α involves an additional phase factor which can be interpreted as polarization. These eight real degrees of freedom are represented by the four complex components of Z^α. Orthogonality $Z^\alpha \bar{Y}_\alpha = 0$ corresponds to collision between the particles. The geometry of twistor space is, in a sense, dual to that of space-time. Zero rest-mass fields (spin $\frac{1}{2}n\hbar$) are described by contour integrals of holomorphic functions of Z^{α} (or of $\bar{Z}_{\alpha}$) which are homogeneous of degree -n-2. Quantum scattering amplitudes for wave packets described by such functions may be evaluated by performing contour integrals in many twistor variables. Each such contour integral may be represented by a graph. Quantum electrodynamics in the high-energy limit can apparently be correctly described in this way, but rest mass and gravitational interactions (which break conformal invariance) have yet to be incorporated correctly. It is hoped that the generalization, suggested by Newman's work, of twistors to asymptotically flat space-time may provide the necessary clue. Asymptotic twistors form a four-complex-dimensional Kähler manifold of signature (++--).

REFERENCES

B. Aronson, R. Lind, J. Messmer and E. Newman, "A note on asymptotically flat spaces" (preprint 1971).

E.T. Newman, E. Couch, K. Chinnapared, A. Exton, A. Prakash and R. Torrence, J. Math. Phys. 6, 918 (1965).

I. Robinson, J. Math. Phys. 2, 290 (1960).

R. Penrose, J. Math. Phys. 8, 345 (1967).

R. Penrose, J. Math. Phys. 10, 38 (1969).

R. Penrose, Int. J. Theoret. Phys. 1, 61 (1968).

R. Penrose, article in a Festschrift dedicated to J.L. Synge (to be published).

R. Penrose, article in a Festschrift dedicated to J.A. Wheeler (to be published).

R. Penrose and M. McCallum, "Twistor theory: an approach to the quantization of fields and of space-time" (preprint 1971).

510 a) 'Radiation, energy and the BMS group'; b) 'Some recent developments'; c) 'N-P constants': 3 summaries of lectures at International Centre for Theoretical Physics (Trieste) Topical meeting (1971) on Gravitation and Field Theory

N-P CONSTANTS

R. Penrose

Birkbeck College, London, UK.

One of the results of the asymptotic analysis of asymptotically flat space-times has been the emergence of a new class of conservation law. For a space-time which is vacuum near $\mathscr{I}^+$ there are five complex (or ten real) N-P constants. Unlike energy and momentum, these quantities remain unaltered when gravitational radiation is emitted by the source. They can be evaluated by examining r^{-6} parts of the curvature. (By comparison, energy-momentum involves r^{-3} parts). If the appropriate correction terms are included, they remain absolutely constant even when electromagnetic or neutrino radiation occurs. These gravitational N-P constants involve a quadratic moment combination of the general form: mass x (quadrupole moment) - (dipole moment)^2, where angular momentum contributions occur as imaginary parts. The N-P constants also involve linear terms which measure what appears, at each retarded time, to be the asymptotic incoming field. For a retarded field these terms arise from back-scattered radiation.

There are also three complex (i.e. six real) N-P constants definable for the asymptotic electromagnetic field (arising from r^{-4} parts of the field). These are constant irrespective of gravitational, electromagnetic or even "charged neutrino" radiation. The stationary contribution is: mass x (electromagnetic dipole) - charge x (gravitational dipole); and again incoming field contributes. There are two complex N-P constants for the neutrino field (treated as a "classical" field). There are also conserved scalar quantities constructable from N-P constants (two complex quantities in the gravitational case and one in the electromagnetic case). N-P constants have not as yet proved very useful in classical relativity. Perhaps they have some significance for quantized theory.

REFERENCES

E.T. Newman and R. Penrose, Phys. Rev. Letters 15, 231 (1965).

E.T. Newman and R. Penrose, Proc. Roy. Soc. (London) A305, 175 (1968).

A. Exton, E.T. Newman and R. Penrose, J. Math. Phys. 10, 1566 (1969).

A. Exton, "Conserved quantities in the Einstein-Maxwell neutrino theory" (Ph.D. Thesis, University of Pittsburgh 1967).

R. Penrose, article in Relativity Theory and Astrophysics , ed. J. Ehlers (Amer. Math. Soc., Providence, RI,1967).

D. Robinson, J. Math. Phys. 10, 1745 (1969).

54

ON A QUADRATIC FIRST INTEGRAL FOR THE CHARGED PARTICLE ORBITS IN THE CHARGED KERR SOLUTION

Comm. Math. Phys. **27(4)** (1972), 303–8. With L. P. Hughston, P. Sommers and M. Walker

See Chapter 43.

Commun. math. Phys. 27, 303–308 (1972)
© by Springer-Verlag 1972

On a Quadratic First Integral for the Charged Particle Orbits in the Charged Kerr Solution*

LANE P. HUGHSTON
Department of Physics: Joseph Henry Laboratories
Princeton University, Princeton, New Jersey, USA
and
Division of Mathematics and Mathematical Physics
The University of Texas at Dallas, Dallas, Texas, USA

ROGER PENROSE
Department of Mathematics, Birkbeck College, London, England

PAUL SOMMERS and MARTIN WALKER**
Center for Relativity Theory, The University of Texas at Austin, Austin, Texas, USA

Received May 2, 1972

Abstract. Associated with the charged Kerr solution of the Einstein gravitational field equation there is a Killing tensor of valence two. The Killing tensor, which is related to the angular momentum of the field source, is shown to yield a quadratic first integral of the equation of the motion for charged test particles.

The Hamilton-Jacobi equation for charged particle orbits in the Kerr spacetime[1] has been found by Carter [1] to be solvable by separation of variables. The separation constant, an expression which is quadratic in the orbital tangent vector, provides in turn a fourth first integral of the equation of motion for charged particles, the other three having been obtained by reason of the two symmetries of the field, and the conservation of the particle rest mass. Walker and Penrose [2] have demonstrated

---

* The research reported herein has been supported in part by the National Science Foundation, grants GP-8868, GP-20023, and GU-1598; the Air Force Office of Scientific Research, grant 903-67; the National Aeronautics and Space Administration, grant 44-004-001; and the Westinghouse Corporation.

** Present address: Max-Planck-Institute for Physics and Astrophysics, 8000 München 40, Föhringer Ring 6, Germany.

[1] Throughout the discussion we refer to the general Kerr family of gravitational fields, with or without either an electromagnetic field [3] or a cosmological constant [1].

in an earlier communication how the fourth first integral of the geodesic equation arises from the existence of a tensor K_{ab} which satisfies[2] $\nabla_{(a}K_{bc)}=0$. In the present communication we show that the quadratic Killing tensor may be used to obtain also Carter's fourth first integral of the equation of motion for *charged test particles.*

To expedite the discussion, we begin with a simplified proof of the existence of a quadratic Killing tensor for the Kerr spacetime. The proof is based on the following remark: *let* o^A *and* ι^A *be a normalized spinor dyad whose corresponding null vectors* $l^a := o^A\bar{o}^{A'}$ *and* $n^a := \iota^A\bar{\iota}^{A'}$ *are tangent to congruences of nonshearing null geodesic curves*[3], *and let* ϕ *be a complex scalar such that* $\phi_{AB}=\phi\, o_{(A}\iota_{B)}$ *satisfies the vacuum Maxwell equation*

$$\nabla^{AA'}\phi_{AB}=0\,. \tag{1}$$

Then the quantity $\mathrm{X}_{AB} := \phi^{-3/2}\phi_{AB}$ *satisfies the twistor equation*[4]

$$\nabla_{(A}{}^{A'}\mathrm{X}_{BC)}=0\,. \tag{2}$$

To see this, we observe that the dyad component

$$6\phi^{3/2}o^A o^B \iota^C \nabla_{(A}{}^{A'}[\phi^{-1/2}o_B\iota_{C)}] = o^A\nabla_A{}^{A'}\phi - 2\phi\, o^B\iota^A\nabla_A{}^{A'}o_B$$

of $\nabla_{(A}{}^{A'}\mathrm{X}_{BC)}$ is simply one of the dyad components of $\nabla^{AA'}[\phi\, o_{(A}\iota_{B)}]$, which vanishes by virtue of the Maxwell equation. The component with o^A and ι^A interchanged vanishes for the same reason, and the remaining two components vanish by virtue of the *GSF* conditions on o^A and ι^A.

The converse of the remark above is also true: If $\nabla_{(A}{}^{A'}\mathrm{X}_{BC)}=0$, then the principal spinors of X_{AB} are *GSF*, and $(\mathrm{X}_{CD}\mathrm{X}^{CD})^{-3/2}\mathrm{X}_{AB}$ satisfies the Maxwell equation. More generally, the principal spinors of a Killing

---

[2] We use the notation of Penrose [4], as does paper [2]. We also make a number of corrections to paper [2]. Eq. (5) should read $\nabla_{(A}{}^{A'}\mathrm{X}_{BC)}=0$. In Eq. (4), the components g_{03} and g_{30} should be $2mraR^{-2}\sin^2\theta$; the component n^2 of Eq. (16) should be $(r^2-a^2-2mr)\,2$, and the component Q^{11} in Eq. (17) should be r^2+a^2-2mr. The equation immediately preceding Eq. (17) should read $Q_{bc}=-2R^2 l_{(b}n_{c)}$. Finally, to be consistent with the *Battelle conventions* for Latin and Gothic indices, the indices in the first line following Eq. (16) should be Latin. The first unnumbered equation after Eq. (17) and the remark following it should read

$$\text{``}\ \delta^{\mathfrak{a}}_a\delta^{\mathfrak{b}}_b\delta^{\mathfrak{c}}_c\nabla^a Q^{bc} = g^{\mathfrak{e}(\mathfrak{a}}\nabla_{\mathfrak{e}}Q^{\mathfrak{bc})} - Q^{\mathfrak{e}(\mathfrak{a}}\nabla_{\mathfrak{e}}g^{\mathfrak{bc})}$$

in any coordinate system $x^{\mathfrak{a}}$, with $\delta^{\mathfrak{a}}_a := \nabla_a x^{\mathfrak{a}}$, so that $\nabla_{\mathfrak{a}} = \partial/\partial x^{\mathfrak{a}}$."

[3] Such spinors will be said to be *GSF*. We notice that the *GSF* condition $o^A o^B\nabla_{AA'}o_B=0$ is preserved under the rescaling $o^A\to\lambda o^A$ for an arbitrary complex scalar λ.

[4] For a conformally flat spacetime, a solution of Eq. (2) describes a symmetric twistor of valence two. A solution of Eq. (2) is generally referred to as a *Killing spinor of valence two.*

spinor of arbitrary valence are *GSF*[5]. In the case of a non-flat vacuum spacetime, this fact would allow us to conclude from the Goldberg-Sachs theorem [5] that at most two of the principal spinors of a Killing spinor of arbitrary valence are not proportional.

Upon examining the Bianchi identities it may be seen readily that in any vacuum spacetime of type {*22*} with a Weyl spinor $\psi o_{(A} o_B \iota_C \iota_{D)}$ we can construct an electromagnetic test field of the kind we have been considering above: $\phi_{AB} = \psi^{2/3} o_{(A} \iota_{B)}$. Moreover, it is well known that in correspondence with each vacuum spacetime of type {*22*} there exists a parametrized family of type {*22*} solutions of the Einstein-Maxwell equations for which the principal null directions of the electromagnetic field coincide with those of the gravitational field[6]. Here the test electromagnetic fields of interest are already at our disposal. Associated, then, with each of these spacetimes is a Killing spinor of valence two[7].

Given the Killing spinor X_{AB} we define the tracefree symmetric tensor

$$P_{ab} := \mathrm{X}_{AB} \bar{\mathrm{X}}_{A'B'} .$$

As a consequence of Eq. (2), P_{ab} is found to satisfy, as it is shown in paper [2], the conformal Killing equation,

$$\nabla_{(a} P_{bc)} = P_{(a} g_{bc)}$$

where g_{ab} is the spacetime metric and where the vector P^a is defined by $P^a := \frac{1}{3} \nabla_b P^{ab}$. If P_a is the gradient of a scalar P, then it may easily be verified that the tensor

$$K_{ab} := P_{ab} - P g_{ab}$$

satisfies the Killing equation:

$$\nabla_{(a} K_{bc)} = 0 . \tag{3}$$

---

[5] Let us consider, for example, an algebraically general Killing spinor $\mathrm{X}_{A_1 \ldots A_n}$ of valence n:

$$\nabla_{A'(A} \mathrm{X}_{A_1 \ldots A_n)} = 0 ,$$

$$\mathrm{X}_{A_1 \ldots A_n} = \overset{1}{\sigma}_{(A_1} \ldots \overset{n}{\sigma}_{A_n)} .$$

Upon transvection n times with $\overset{n}{\sigma}_A$ the twistor equation becomes

$$\overset{n}{\sigma}{}^A \overset{n}{\sigma}{}^{A_1} \ldots \overset{n}{\sigma}{}^{A_n} \nabla_{A'(A} \mathrm{X}_{A_1 \ldots A_n)}$$

$$\left(\overset{n}{\sigma}{}^{A_1} \overset{1}{\sigma}_{A_1}\right) \ldots \left(\overset{n}{\sigma}{}^{A_{n-1}} \overset{n-1}{\sigma}{}_{A_{n-1}}\right) \left| \overset{n}{\sigma}{}^A \overset{n}{\sigma}{}^{A_n} \nabla_{A'A} \overset{n}{\sigma}_{A_n} \right| = 0$$

which is seen to be the *GSF* condition on the congruence defined by $\overset{n}{\sigma}_A$.

[6] Kinnersley [6] has provided an explicit algorithm for generating the electrification of any type {*22*} vacuum spacetime.

[7] The argument given in paper [2] for the existence of a Killing spinor of valence 2 is not quite complete as it stands for the *charged* Kerr solution, since the proof given would seem to rely on the vacuum Bianchi identities. Our procedure here should clarify that derivation.

Let us examine the condition that P_a be a gradient. This condition may be expressed in a particularly convenient form if we first observe that the stress-energy tensor

$$\tau_{ab} := \phi_{AB}\bar{\phi}_{A'B'} = \phi\bar{\phi}[l_{(a}n_{b)} - \tfrac{1}{4}g_{ab}]$$

of the test electromagnetic field ϕ_{AB} is proportional to the conformal Killing tensor P_{ab}, and satisfies

$$\nabla^a \tau_{ab} = 0$$

as a consequence of the vacuum Maxwell equation. From these facts it follows that

$$P_a = \tfrac{1}{3}\nabla^b P_{ab} = f^2 \tau_{ab} \nabla^b f = l_{(a}n_{b)} \nabla^b f - \tfrac{1}{4}\nabla_a f$$

where $f := (\phi\bar{\phi})^{-1/2}$. Therefore, the requirement that P_a be a gradient is equivalent to the existence of a scalar h such that

$$l_{(a}n_{b)} \nabla^b f = \nabla_a h .$$

We shall now consider explicitly the case of the Kerr spacetime. In the coordinate system of paper [2], the scalar ϕ is given by

$$\phi = (r - ia\cos\theta)^{-2}$$

so that $f = r^2 + a^2\cos^2\theta$. A simple computation shows that

$$l_{(a}n_{b)} \nabla^b f = \nabla_a(\tfrac{1}{2}r^2) .$$

Thus we establish that *the Kerr spacetime admits a Killing tensor of valence two*[8].

The equation of motion of a test particle with charge ε, moving under the influence of an electromagnetic field F_{ab}, may be written

$$t^c \nabla_c t^a = \varepsilon t^c F^a{}_c , \qquad (4)$$

where t^a is tangent to the orbit of the particle. As in the case of an affinely parametrized geodesic orbit, as would be given by $\varepsilon = 0$, the equation of motion admits $g^{ab}t_a t_b$ as a first integral since

$$t^c \nabla_c (g^{ab} t_a t_b) = 2t^b t^c \nabla_c t_b = 2\varepsilon t^b t^c F_{bc} = 0 .$$

Suppose that g_{ab} admits a one parameter group of isometries generated by a Killing vector ξ^a, so that $\mathscr{L}_\xi g_{ab} = 0$. A first integral may be constructed from ξ^a provided that the electromagnetic field is also invariant under the action of the group. In fact, we may demonstrate that $\mathscr{L}_\xi F_{ab} = 0$ is the necessary and sufficient condition that there exist a

---

[8] We comment, incidentally, that our analysis is more generally valid for the charged *Kerr-NUT* spacetimes [1, 7] wherein the Killing tensor assumes the form

$$K_{ab} = (l - a\cos\theta)^2\, l_{(a}n_{b)} + r^2 m_{(a}\bar{m}_{b)}$$

where l is the *NUT* parameter.

scalar α such that

$$t^c \nabla_c(\xi^a t_a + \varepsilon\alpha) = 0\,. \tag{5}$$

Using the equation of motion (4) we find Eq. (5) to be satisfied if and only if

$$\nabla_a \alpha + F_a{}^c \xi_c = 0$$

for which we have the integrability condition

$$\nabla_{[a}(F_{b]c}\xi^c) = 0\,.$$

Then, using Maxwell's equations in the form

$$2\nabla_{[a}F_{b]c} + \nabla_c F_{ab} = 0$$

we arrive at the desired relation:

$$\xi^c \nabla_c F_{ab} + F_{ac}\nabla_b \xi^c + F_{cb}\nabla_a \xi^c := \mathscr{L}_\xi F_{ab} = 0\,.$$

If a local vector potential A_a for F_{ab} be chosen with the gauge condition $\mathscr{L}_\xi A_a = 0$, then $\alpha = \xi^a A_a$, and the conserved quantity is given by

$$\xi^a(t_a + \varepsilon A_a)\,.$$

Thus far our discussion of first integrals has been valid in general. We restrict our attention now to those spacetimes which admit a Killing spinor X_{AB}, and ask under what conditions we can construct from X_{AB} a quadratic first integral for the orbits defined by Eq. (4). First we consider the case in which the orbit is *null*[8]. If the tangent to the orbit is written $t^a := \lambda^A \bar{\lambda}^{A'}$, then the equation of motion (4) assumes the form

$$\begin{gathered}\bar{\lambda}_{A'}(t_c \nabla^c \lambda_A + \varepsilon\phi_{AB}\lambda^B) + \lambda_A(t_c \nabla^c \bar{\lambda}_{A'} + \varepsilon\bar{\phi}_{A'B'}\bar{\lambda}^{B'}) = 0 \\ F_{ab} := \phi_{AB}\varepsilon_{A'B'} + \bar{\phi}_{A'B'}\varepsilon_{AB}\end{gathered} \tag{6}$$

from which we deduce that

$$t^c \nabla_c \lambda_A + \varepsilon\phi_{AB}\lambda^B = i\omega\lambda_A$$

for some real scalar ω. A suitable new choice for the phase of λ_A serves to eliminate ω. Then the equation of motion becomes

$$t^c \nabla_c \lambda_A + \varepsilon\phi_{AB}\lambda^B = 0\,.$$

It is easy to see that the complex scalar $\mathsf{X} := \mathsf{X}_{AB}\lambda^A\lambda^B$ formed from the Killing spinor X_{AB} is constant along the orbits, $t^a \nabla_a \mathsf{X} = 0$, provided that

$$\mathsf{X}_{B(A}\phi_{C)}{}^B = 0\,,$$

which is precisely the condition that X_{AB} and ϕ_{AB} be proportional. Therefore, *for those spacetimes in which the principal null directions of the electromagnetic field are aligned with those of the Killing spinor*

---

[9] The null charged particle orbits are, due to their conformal invariance, of some interest in connection with the theory of *twistor quantization* [8].

we obtain a first integral for the charged null orbits. This condition is fulfilled, of course, for the charged Kerr spacetime.

We now demonstrate the construction of a first integral from the Killing tensor

$$K_{ab} := \mathrm{X}_{AB}\bar{\mathrm{X}}_{A'B'} - P g_{ab}$$

for the non-null charged particle orbits in the Kerr case. Applying Eqs. (3), (4), and (6) we find that

$$t^c \nabla_c (K^{ab} t_a t_b) = 2\varepsilon t^a t^c K_{b(a} F_{c)}{}^b = -2\varepsilon t^a t^c [\mathrm{X}_{AC} \bar{\mathrm{X}}_{B'(A'} \bar{\phi}^{B'}{}_{C')} + \bar{\mathrm{X}}_{A'C'} \mathrm{X}_{B(A} \phi^B{}_{C)}] ,$$

which vanishes identically due to the proportionality $\mathrm{X}_{AB} = \phi^{-3/2} \phi_{AB}$. Thus we obtain the first integral

$$K^{ab} t_a t_b .$$

We remark, finally, on the *physical* meaning of the Killing spinor in Kerr's spacetime. Suppose that from the Killing spinor X_{AB} we construct the bivector

$$M_{ab} := \mathrm{X}_{AB} \varepsilon_{A'B'} + \bar{\mathrm{X}}_{A'B'} \varepsilon_{AB} .$$

Then in the linear approximation to Einstein's theory, this bivector can be shown to be related to the expression for the total angular momentum of the source of the gravitational field[10].

The authors are grateful to Jürgen Ehlers and Ivor Robinson for their conversation.

References

1. Carter, B.: Commun. math. Phys. **10**, 280 (1968); — Phys. Rev. **174**, 1559 (1968).
2. Walker, M., Penrose, R.: Commun. math. Phys. **18**, 265 (1970).
3. Newman, E. T., Couch, E., Chinnapared, R., Exton, A., Prakash, A., Torrence, R.: J. Math. Phys. **6**, 918 (1965).
4. Penrose, R.: Structure of space-time. In: De Witt, C., Wheeler, J. A. (Eds.): Battelle Rencontres. New York: W. A. Benjamin Inc. 1968.
5. Goldberg, J., Sachs, R. K.: Acta Phys. Polon. **22**, 13 (1962).
6. Kinnersley, W. M.: Type D gravitational yields (unpublished thesis, California Institute of Technology), (1969).
7. Miller, J. G.: Global analysis of the Kerr-Taub-NUT metric (Princeton University preprint) (1972).
8. Penrose, R., MacCallum, M. A. H.: Twistor theory: An approach to the quantization of fields and space-time (Notes on lectures delivered at the Institute of Theoretical Astronomy, Cambridge) (1972).

Prof. R. Penrose
Department of Mathematics
Birkbeck College
London, England

---

[10] This last remark can be made in a more concise way in the language of twistors, to which, in the linearized theory, Killing spinors have an especially close relation.

55

THE GEOMETRY OF IMPULSIVE GRAVITATIONAL WAVES

General relativity (papers in honour of J. L. Synge). Ed. L. O'Raifeartaigh, pp. 101–15. Clarendon Press, Oxford, 1972.

The impulsive gravitational waves described here (where the curvature is a delta function) already featured in Chapters 37 and 52. Various types of intrinsic geometry for null 3-surfaces in space-time are also proposed, 2-spinor methods being used.

The geometry of impulsive gravitational waves. *General relativity (papers in honour of J. L. Synge)*. Ed. L. O'Raifeartaigh, pp. 101–15

THE GEOMETRY OF IMPULSIVE GRAVITATIONAL WAVES

ROGER PENROSE

ONE of the many topics pioneered by Professor Synge is the study of space–times containing *discontinuities* in the curvature.[1],[2],[3],[4],[5] For such a space–time M, the metric is taken to be C^1 everywhere, failing to be (at least) C^3 only on some hypersurface N—in accordance with the Lichnerowicz conditions.[6] A more extreme situation has also been studied, in which the metric of the space–time M is taken merely to be C^0 everywhere, but again failing to be C^3 only on a hypersurface N.[7],[8],[9],[10] In such cases there will be a δ-function in the curvature along N. When M is an empty space–time (or, at least, where the Ricci tensor remains continuous), then in the first case we may refer to the curvature discontinuity as a gravitational *shock wave*, while in the second case we may refer to the curvature δ-function as a gravitational *impulsive wave*. The distinction between these two situations is illustrated in Fig. 1. A two-dimensional closed surface is depicted that consists of a hemisphere, a portion of a cylinder, and a disc, joined together along two circles. The Riemannian metric is C^∞ except at the two circles, being C^1 (discontinuity in curvature) at the junction of the hemisphere with the cylinder and C^0 (δ-function in curvature) at the junction of the cylinder

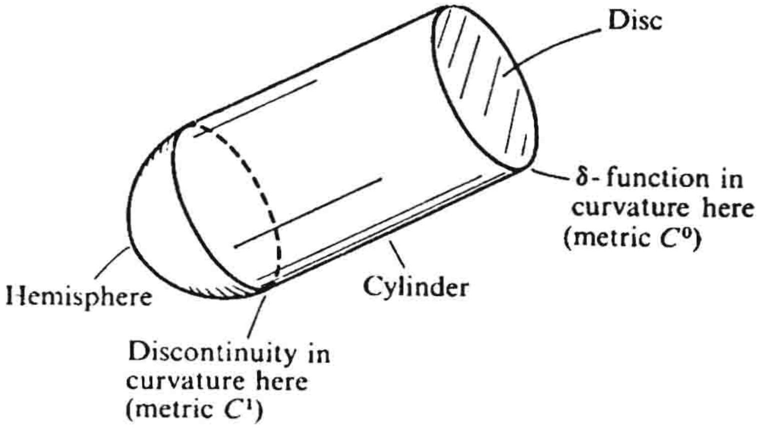

Disc
δ-function in curvature here (metric C^0)
Hemisphere
Cylinder
Discontinuity in curvature here (metric C^1)

FIG. 1. A two-dimensional closed surface exhibiting both a discontinuity and a δ-function in the curvature.

with the disc. This illustration also serves to indicate something of the geometrical attitude favoured by Professor Synge that I wish to adopt in this article.

The scissors-and-paste approach that I am adopting is similar in some respects to that of Israel.[11] However, I shall be using spinor techniques here to aid the geometry. This will enable simple explicit formulae to be given that relate the discontinuity in the curvature (and its Petrov type) across N to the nature of the δ-function along N and to the geometry of N itself. Comparisons will be made with Maxwell theory (yielding a valid form of 'Hadamard's conjecture' for Maxwell theory, in contrast with the negative result of Günther for the scalar case).[12],[13],[14] I shall also be led to consider various possible concepts of intrinsic geometry for null hypersurfaces. This has some interest also in connection with the characteristic initial-value problem.[15],[16],[17],[18]

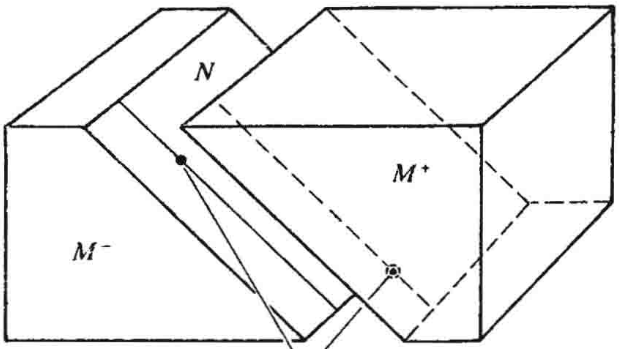

N
M⁺
M⁻
Identify in warped fashion by shunting along null generators

FIG. 2. A plane-fronted impulsive wave can be constructed by joining two Minkowski half-spaces together with a warp along a null hyperplane.

Let me begin by giving two examples of vacuum space–times that represent pure impulsive gravitational waves—that is to say, M is completely flat except along the hypersurface N (compare the cylinder–disc part of the surface of Fig. 1). In fact, as I shall indicate later, these two examples will represent the only possible space–times of this type.

For the first example,[19],[20] we choose coordinates for Minkowski space–time so that the metric takes the form

$$ds^2 = 2\mathrm{d}u\mathrm{d}v - 2\mathrm{d}\zeta\mathrm{d}\bar{\zeta} \tag{1}$$

where u and v are real, ζ is complex. Removal of the null hyperplane $v = 0$ divides the space into two halves, say M^+ (given by $v > 0$) and M^- (given by $v < 0$). Now, we re-attach a null hyperplane N (given by $v = 0$) as the common boundary to each of M^- and M^+, where we make the identification

$$(u,0,\zeta,\bar{\zeta})_{M^-} \equiv (u-f(\zeta,\bar{\zeta}),0,\zeta,\bar{\zeta})_{M^+} \tag{2}$$

between the two halves (see Fig. 2), f being a function of ζ and $\bar{\zeta}$ alone. Einstein's vacuum equations for the resulting space–time turn out to be

$$\frac{\partial^2 f}{\partial\zeta\partial\bar{\zeta}} = 0, \tag{3}$$

which is a consequence of later results. Condition (3) can be expressed as

$$f(\zeta,\bar{\zeta}) = g(\zeta)+\bar{g}(\bar{\zeta}),$$

where g is an analytic (holomorphic) function.

The metric for this space–time can in fact be written

$$ds^2 = 2dv\{du - f(\zeta,\bar{\zeta})\delta(v)dv\} - 2d\zeta d\bar{\zeta}, \tag{4}$$

which is an idealized plane-fronted wave form.[21] Note that a δ-function appears in the metric components. This is due to the fact that the coordinates, being chosen so that the Minkowski form (1) holds in each of M^- and M^+, are discontinuous across N. The metric tensor field itself (regarded in the modern coordinate-free way) is actually C^0. We can see this by transforming to a new coordinate system that is continuous[22],[23],[24] (but in many ways less convenient than the one used in eqn (4)). Let us consider explicitly the special case $f \equiv \frac{1}{2}(\zeta^2+\bar{\zeta}^2)$, which gives a plane wave. The discontinuous coordinate transformation to continuous real coordinates (U,V,X,Y) given by

$$\begin{aligned}\sqrt{2}\zeta &= X(1+V\theta)+iY(1-V\theta),\\ u &= U+\tfrac{1}{2}X^2(V+1)\theta+\tfrac{1}{2}Y^2(V-1)\theta,\\ v &= V,\end{aligned}$$

where $\theta = \theta(V)$ is Heaviside's step function, reduces the metric form (4) to:

$$ds^2 = 2dUdV-(1+V\theta)^2dX^2-(1-V\theta)^2dY^2;$$

for which the metric components are now seen to be continuous. The jump in the $(u,v,\zeta,\bar{\zeta})$ coordinates at $v = 0$ is simply because u, at $v = 0-$, becomes $u-\frac{1}{2}(\zeta^2+\bar{\zeta}^2)$, at $v = 0+$, as required for eqn (2). The general metric form (4) can be derived in a similar way, but the explicit expressions tend to get rather complicated.

The second example of a pure impulsive gravitational wave is actually a limiting case of a Robinson–Trautman null spherically fronted wave.[25] We choose new coordinates for Minkowski space–time (derived from our earlier ones by the transformations $u \to u$, $v \to v-\zeta\bar{\zeta}u$, $\zeta \to u\zeta$), so that the metric takes the form

$$ds^2 = 2dudv-2u^2d\zeta d\bar{\zeta}. \tag{5}$$

Now, $v = 0$ is a *null cone*, the removal of which leaves disjoint pieces M^- (given by $v < 0$) and M^+ (given by $v > 0$). We re-attach M^- to M^+ along a null hypersurface N, identifying according to the relation

$$(u,0,\zeta,\bar{\zeta})_{M^-} = \left(\frac{u}{|h'(\zeta)|}, 0, h(\zeta), h(\bar{\zeta})\right)_{M^+}, \tag{6}$$

where h is analytic.† It turns out that now Einstein's equations are automatically satisfied.

As in the previous example, the (abstract) metric tensor is C^0. This can in fact be seen by a transformation to a continuous coordinate system, but this is complicated. Moreover, a new situation does arise in this case, since if the discontinuous $(u,v,\zeta,\bar{\zeta})$ system is used, the metric components in this system involve squares of δ-functions, which is not really allowable. But nevertheless, it is unnecessary to go to the trouble of actually constructing a continuous coordinate system for this example in order to see that its metric is C^0. We merely have to ascertain[11] that the intrinsic metric of N as induced from M^- agrees with that induced from M^+. This is readily checked by substitution of eqn (6) into eqn (5).

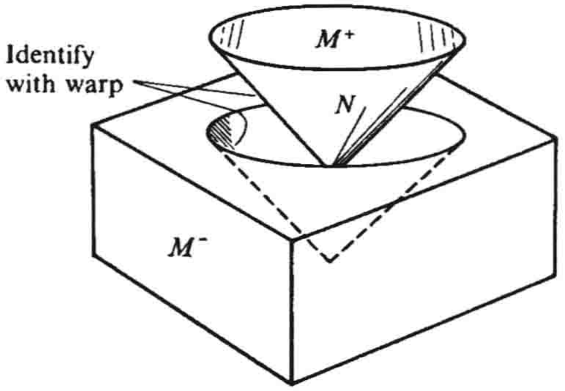


FIG. 3. A Robinson–Trautman impulsive null gravitational wave, constructed by joining the interior of a light cone to its exterior with a suitable warp.

Before showing how Einstein's equations are satisfied for examples constructed in such a manner, I should say a little more about the nature of intrinsic geometry for a general null hypersurface N. The geometry of such hypersurfaces is not quite a straightforward matter. It is possible to distinguish at least three distinct types of geometry for N, all of which can be reasonably referred to as intrinsic. This difficulty stems from the fact that, since the induced metric for N is degenerate, it is not possible to define a unique connection from this metric.

---

† The singularities in h imply the existence of 'wire singularities' along N. I shall not concern myself with such matters here. All the considerations of this article will be essentially local.

We recall that any null hypersurface N in a space–time M is generated by a uniquely defined two-parameter family of null geodesics (γ), one through each point of N. Now N inherits from M:

(a) a degenerate three-metric;
(b) a concept of affine parameter on each γ or, equivalently, a concept of parallel transport along γ of tangent vectors to γ;
(c) a concept of parallel transport, along each γ, of tangent vectors to N at points of γ.

All of these concepts could reasonably be considered as intrinsic to N. (Note that in (b) we do not single out any *particular* affine parameter for each γ—this could not really be intrinsic, since it is not even defined naturally by the geometry of M.) The three types of intrinsic geometry are, in order of increasing structure:

I-geometry—concept (a) only;
II-geometry—concepts (a) and (b);
III-geometry—concepts (a) and (c).

In order to study these different geometries in detail, we set up the following framework. At each point of N we choose a two-spinor o^A, smoothly, such that the null vector constructed from it,†

$$l^a = o^A \bar{o}^{A'},$$

points in the null direction tangent to N (that is, in the direction of γ, i.e. normal to N). We have the gauge freedom:

$$o^A \to \lambda o^A, \tag{7}$$

where λ is a nowhere-vanishing smooth scalar function on N. This rescales l^a according to $l^a \to |\lambda|^2 l^a$, but leaves the direction of l^a unchanged. In addition,[19] the 'flag plane' defined by o^A (a null two-plane element containing l^a) is rotated through an angle 2 arg λ under eqn (7). We shall be concerned with weighted quantities $\eta^{\cdots}_{\cdots}$ which depend on o^A to the extent of

$$\eta^{\cdots}_{\cdots} \to \lambda^p \bar{\lambda}^q \eta^{\cdots}_{\cdots} \tag{8}$$

whenever o^A undergoes (7). Such an $\eta^{\cdots}_{\cdots}$ is said to have *type* (p,q).[26],[17]

Now suppose N to be imbedded in a space–time in two alternative ways. We may conveniently imagine both imbeddings to be carried out simultaneously if we have in mind a picture such as that considered earlier. In this case, N is imbedded as the common boundary to two half space–times M^- and M^+. The spinor field o^A is continued smoothly from N into $M^-$and also smoothly from N into M^+. (We cannot expect that both of these continuations will fit smoothly on to each other, however.) We require that the intrinsic metric of N (i.e. the I-geometry) be the same when induced by that of M^- as by that of M^+ (otherwise we should have difficulty assigning a meaning to the concept of a spinor† in N).

---

† I am adopting Battelle conventions here,[19] whereby tensor and spinor indices are treated as (non-numerical) abstract labels. Infeld–van der Waerden translation symbols between tensor and spinor indices are therefore unnecessary.

We denote the covariant derivative in M^- by $\nabla_a^- = \nabla_{AA'}^-$ and in M^+ by $\nabla_a^+ = \nabla_{AA'}^+$. Since any vector tangent to N has, for some ξ^A, a spinor form $\xi^A\bar{o}^{A'}+o^A\bar{\xi}^{A'}$ (being orthogonal to $l^a = o^A\bar{o}^{A'}$), it follows that the derivatives acting within N are spanned by

$$\bar{o}^{A'}\nabla_{AA'} \quad \text{and} \quad o^A\nabla_{AA'}.$$

Here $\nabla_{AA'}$ stands for either $\nabla_{AA'}^+$ or $\nabla_{AA'}^-$. When acting on scalars defined on N, each operator yields the same result whichever of $\nabla_{AA'}^+$ or $\nabla_{AA'}^-$ is used. But when acting on a spinor or tensor defined on N there will in general be a difference.

This jump in the connection on M corresponds to the δ-function in the curvature at N. We write the Riemann tensor in a standard spinor form[19]

$$R_{abcd} = \mathrm{X}_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'}+\varepsilon_{AB}\varepsilon_{CD}\bar{\mathrm{X}}_{A'B'C'D'}+ \\ +\varepsilon_{AB}\Phi_{CDA'B'}\varepsilon_{C'D'}+\varepsilon_{A'B'}\Phi_{ABC'D'}\varepsilon_{CD}, \tag{9}$$

with

$$\Psi_{ABCD} = \mathrm{X}_{(ABCD)}, \quad \Lambda = \tfrac{1}{3}\mathrm{X}_{AB}{}^{AB}. \tag{10}$$

The spinor Ricci identities are

$$\nabla_{X(A'}\nabla_{B')}{}^{X}\xi_C = \Phi_{CDA'B'}\xi^D, \quad \nabla_{Y'(A}\nabla_{B)}{}^{Y'}\xi_C = \mathrm{X}_{ABCD}\xi^D. \tag{11}$$

Writing $\nabla_{AA'} = \Theta\nabla_{AA'}^+ + (1-\Theta)\nabla_{AA'}^-$, where Θ is the step function taking the value 0 on M^- and 1 on M^+, and defining the generalized function Δ by

$$\nabla_{AA'}\Theta = o_A\bar{o}_{A'}\Delta, \tag{12}$$

we obtain, as the δ-function part of eqn (11),

$$o^A(\nabla_{AA'}^+ - \nabla_{AA'}^-)\xi_B = -o_{(B}\phi_{C)A'}\xi^C \tag{13}$$

and

$$\bar{o}^{A'}(\nabla_{AA'}^+ - \nabla_{AA'}^-)\xi_B = -o_{(B}\psi_{C)A}\xi^C \tag{14}$$

on N. Here $\phi_{CA'}(=\bar{\phi}_{A'C})$ and $\psi_{CA}(=\psi_{AC})$ measure the δ-function parts of the curvature (on N) according to

$$\overset{\delta}{\mathrm{X}}_{ABCD} = o_{(A}\psi_{B)(C}o_{D)}\Delta, \quad \overset{\delta}{\Phi}_{ABC'D'} = o_{(A}\phi_{B)(C'}\bar{o}_{D')}\Delta. \tag{15}$$

---

† Actually, one could proceed in this way if the intrinsic *conformal* metrics only were matched.[15] But I shall not be concerned with such more general situations here.

From eqn (10) we obtain the δ-function in the Weyl curvature and scalar curvature respectively:

$$\overset{\delta}{\Psi}_{ABCD} = \psi_{(AB}o_C o_{D)}\Delta, \quad \overset{\delta}{\Lambda} = -\tfrac{1}{6}\psi_{AB}o^A o^B \Delta. \tag{16}$$

Now the most general situation occurs when it is merely the I-geometry of N induced by M^-, which agrees with that induced by M^+. Then the real vector $\phi_a = \phi_{AA'}$ defined on N is, in itself, pointwise unrestricted.† The symmetric spinor ψ_{AB} is restricted pointwise only in that $\psi_{AB}o^A o^B$ must be *real* and in that ψ_{AB} and $\phi_{AA'}$ are connected by the equation

$$\psi_{AB}o^B = \phi_{AB'}\bar{o}^{B'}, \tag{17}$$

which follows from eqns (13) and (14) by transvection with $\bar{o}^{A'}\xi^B$ and $o^A\xi^B$ respectively. Note that even in this general case, when no field equations are imposed, it follows that the algebraic (i.e. Petrov) type of the δ-function part of the Weyl tensor is special.[19],[27] From eqn (16) we see that its type must be {211} or more special, with the repeated principal null direction pointing along the null direction in N. Furthermore, the infinite part of the scalar $J = \Psi_{ABCD}\Psi^{ABCD}$ has to be real.

Each generator γ of N is a null geodesic, so $o^A\bar{o}^{A'}\nabla^{\pm}_{AA'}o_B = l^a\nabla^{\pm}_a o_B = \varepsilon^{\pm}o_B$ for some ε^+, ε^-. By eqn (13) (or eqn (14)), it follows that $I_m(\varepsilon^+) = I_m(\varepsilon^-)$. We can, if desired, choose o_B so that $\varepsilon^- = 0$. Then ε^+ must be real. Thus, o_B will be parallelly propagated along γ according to ∇^-_a, but only the flag plane[19] of o_B (and not its extent) is parallelly propagated along γ according to ∇^+_a. Thus, the concept of parallel propagation along γ of null two-plane elements tangent to N (i.e. flag planes) is invariantly defined by the I-geometry.

Next, suppose that the II-geometry of N induced by M^- agrees with that induced by M^+. This means that the concept of parallel propagation of the extent of o_B (i.e. of the null vector l^a) is also invariantly defined. Thus we must have $\varepsilon^- = \varepsilon^+$, which means that

$$\psi_{AB}o^A o^B = 0 = \phi_{AB'}o^A\bar{o}^{B'}. \tag{18}$$

That is to say, the δ-function part of the scalar curvature vanishes. Note that eqn (18) implies that the vector ϕ_a is now tangent to N; also that $\psi_{AB} = \alpha_{(A}o_{B)}$ for some α_A. Thus, the δ-function in the Weyl tensor is now type {31} or more special.

Finally, let us consider III-geometry. Any tangent vector to N has the form $\xi^A\bar{o}^{A'} + o^A\bar{\xi}^{A'}$, so a concept of parallel propagation of tangent vectors to N, along generators of N, will be defined if we can define a concept of parallel propagation of spinors ξ^A along each generator γ. Conversely, the propagation of ξ^A follows from the propagation of tangent vectors to N, since (if $\xi^A o_A \neq 0$) the two orthogonal tangent vectors $\xi^A \bar{o}^{A'} + o^A \bar{\xi}^{A'}$ and $\mathrm{i}\xi^A \bar{o}^{A'} - \mathrm{i}o^A \bar{\xi}^{A'}$ together (with o^A) define ξ^A. Thus, the III-geometry induced by M^- agrees with that induced by M^+ if and only if $l^a \nabla^-_a \xi_B = l^a \nabla^+_a \xi_B$ for all ξ_B, i.e. (from eqns (13) and (14))

---

† If we are interested in impulsive waves of null matter, satisfying an energy condition such as non-negative definiteness of its energy density, then ϕ_a must be taken as non-space-like and future-pointing.

$$\psi_{AB} o^B = 0 = \phi_{AB'} \bar{o}^{B'}. \tag{19}$$

Using eqn (19), we can write

$$\psi_{AB} = \psi o_A o_B, \quad \phi_{AB'} = \phi o_A \bar{o}_{B'} \tag{20}$$

(where ϕ is real), whence

$$\overset{\delta}{\Psi}_{ABCD} = \psi o_A o_B o_C o_D \Delta, \quad \overset{\delta}{\Phi}_{ABC'D'} = \phi o_A o_B \bar{o}_{C'} \bar{o}_{D'} \Delta, \quad \overset{\delta}{\Lambda} = 0. \tag{21}$$

Note that the δ-function in the Weyl tensor is now type {4} (or {−}).

Let us examine the effect of imposing Einstein's equations on the impulsive wave. For the wave to be purely gravitational—whether or not the wave is passing through matter, and independently of a λ-term in the equations—we must have no δ-function in the trace-free Ricci tensor. Thus $\phi_{AB} = 0$ and, by eqn (17), $\psi_{AB} o^B = 0$. The condition (19) is satisfied, so the III-geometry induced by M^- agrees with that induced by M^+. The δ-function in the Weyl tensor is therefore type {4}.[7],[8]

There is a simple geometric test for deciding whether or not an impulsive wave is purely gravitational. Consider a null hypersurface H^- in M^- intersecting N in a space-like two-surface S. There will be a unique continuation H^+ of H^- into M^+, where H^+ is the null hypersurface in M^+ meeting N in S. Let ξ^A point along the null generators of $M^\pm$, being chosen to be continuous, with $\xi_A o^A = 1$ at N, and smooth on $H^\pm$. Then the quantities

$$\bar{\xi}^{A'} \xi^B o^A \nabla^\pm_{AA'} \xi_B = \mathrm{P}^\pm \quad \text{and} \quad \xi^A \xi^B \bar{o}^{A'} \nabla^\pm_{AA'} \xi_B = \Sigma^\pm \tag{22}$$

describe,[19] respectively, the *convergence* and *shear* of $H^\pm$ at S. From eqns (13) and (14) we get

$$\mathrm{P}^+ - \mathrm{P}^- = \phi_{CA'} \xi^C \bar{\xi}^{A'}, \quad \Sigma^+ - \Sigma^- = \psi_{CA} \xi^C \xi^A.$$

Thus, a purely gravitational impulse wave occurs if the convergence of $H^\pm$ is necessarily continuous, the strength of the wave being then measured by the jump in the shear.[28] For this we need choose only one null hypersurface corresponding to each null direction at each point of N. Thus, we can test our original examples, eqns (2) and (6), simply by choosing H^- to be a general null hyperplane in M^-. It is then a straightforward matter to check the vanishing of the convergence (at N) of the emergent null hypersurface H^+ for eqn (6) and for eqns (2) and (3), hence verifying Einstein's vacuum equations for these cases.

Let us investigate the various null-hypersurface geometries in a little more detail. It will be convenient to introduce certain differential operators $\eth_{A'B}$, $\bar{\eth}_{BA'}$, þ, acting within N, which send weighted quantities (eqn (8)) into other such quantities.† Let $\eta^{\cdots}_{\cdots}$ be a weighted quantity of type (p,q). Then $\eth_{AB'}\eta^{\cdots}_{\cdots}$ will have type $(p+2, q)$, being defined by the equation

$$\bar{o}_{(D'}\eth_{A')B}\eta^{\cdots}_{\cdots} = o_B o^C \bar{o}_{(D'}\nabla_{A')C}\eta^{\cdots}_{\cdots} - \eta^{\cdots}_{\cdots}(p o^C \bar{o}_{(D'}\nabla_{A')C} o_B + q o_B o^C \nabla_{C(A'}\bar{o}_{D')}), \qquad (23)$$

giving two versions $\eth^-_{A'B}$ and $\eth^+_{A'B}$ of $\eth_{A'B}$ according as ∇^-_a or ∇^+_a is used on the right. We readily verify that the right-hand side of eqn (23) is in fact a weighted quantity of type $(p+2, q+1)$ (compare eqns (7), (8)) and furthermore that it has the required form $\bar{o}_{(D'}\zeta_{A')B}{}^{\cdots}_{\cdots}$ since transvection with $\bar{o}^{D'}\bar{o}^{A'}$ gives zero (because $\bar{o}^{D'}l^a\nabla_a\bar{o}_{D'} = 0$). It follows that each of $\eth^-_{A'B}$ and $\eth^+_{A'B}$ is uniquely defined by eqn (23) as a weighted operator of type (2,0). The operator $\bar{\eth}_{BA'}$ is similarly defined, as the complex conjugate of $\eth_{B'A}$, and is a weighted operator of type (0,2). The operator $\text{þ}(=\bar{\text{þ}})$ of differentiation of propagation along generators of N is a weighted (scalar) operator of type (1,1) defined by

$$\text{þ}\eta^{\cdots}_{\cdots} = l^a\nabla_a\eta^{\cdots}_{\cdots} - (p\varepsilon + q\bar{\varepsilon})\eta^{\cdots}_{\cdots}, \qquad (24)$$

where $l^a\nabla_a o_B = \varepsilon o_B$. Expressions (23) and (24) are simpler when o_A is scaled so that $\varepsilon = 0$. But we cannot arrange both $\varepsilon^- = 0$ and $\varepsilon^+ = 0$ in the most general case. Then we have distinct operators þ^- and þ^+.

When acting on weighted scalars, the operators $\eth_{A'B}$ and þ are, respectively, invariants of III-geometry and II-geometry (and both invariants of I-geometry if $p = -q$), in the sense that $\eth^+_{A'B}\eta = \eth^-_{A'B}\eta; \text{þ}^+\eta = \text{þ}^-\eta$. When acting on a general weighted spinor, þ is an invariant of III-geometry, but $\eth_{AB'}$ is not. (We can consider a richer geometry yet, namely the IV-geometry, with structure defined by the operator $\eth_{A'B}$. But this could less reasonably be called intrinsic to N.) Certain expressions constructed from $\eth_{A'B}$ are, however, invariants of I, II, or III-geometry. In every case we can check the invariance by referring back to eqns (13) and (14), and using eqn (17), (18), or (19) in the case of I, II, or III-geometry respectively. However, it is easier to employ the following explicit formulae for the jumps in $\eth_{A'B}$ and þ. We denote the jump in a quantity across N by square brackets [ ] (thus $[\text{þ}\eta] = \text{þ}^+\eta - \text{þ}^-\eta$, etc.) and write

$$\chi_A = \phi_{AB'}\bar{o}^{B'} = \psi_{AB}o^B.$$

Let $\eta^{\cdots}_{\cdots}$ have type (p,q) and put $r = \frac{1}{2}(p+q)$. In the case of a weighted scalar we have

$$[\eth_{A'B}\eta] = r\eta o_B\bar{\chi}_{A'}, \quad [\text{þ}\eta] = r\eta o^A\chi_A;$$

---

† If ι^A is introduced, with $\iota^A o_A = 1$, so as to complete a spinor dyad, then $\iota^A\bar{\iota}^{B'}\eth_{B'A}$ becomes the operator $\eth$ of Geroch, Held, and Penrose[26] which in turn generalizes the operator $\eth$ of Newman and Penrose.[29] The þ operator used here agrees with that of ref. 26.

for a weighted spin-vector

$$[\eth_{A'B}\eta_C] = -\eta^D o_{(C}\phi_{D)A'}o_B + r\eta_C o_B\bar{\chi}_{A'}, \quad [\text{þ}\eta_A] = -\eta^B o_{(A}\chi_{B)} + r\eta_A o^B\chi_B,$$

or

$$[\eth_{A'B}\eta_{C'}] = -\eta^{D'}\bar{o}_{(C'}\bar{\psi}_{D')A'}o_B + r\eta_{C'}o_B\bar{\chi}_{A'}.$$

It should be remarked, here, that any weighted spinor quantity defined on N has a geometric meaning intrinsic to N. For example, if ξ^A has type $(-1,0)$, then the normalizing factor $\xi^A o_A$ is an ordinary scalar (type $(0,0)$), the remaining information of ξ^A defining a two-plane element tangent to N (determined by orthogonality with the flag-pole direction of ξ^A). The interpretation of a general weighted $\eta^{\cdots}_{\cdots}$ proceeds along similar lines. A weighted scalar η defines null two-plane elements tangent to N corresponding to positions of the flag plane of o^A which makes η positive; and so on.

Let me give some examples of invariant weighted quantities. We have

$$\eth_{A'B}o_C = -\sigma\bar{o}_{A'}\varepsilon_{BC}, \quad \bar{\eth}_{AB'}o_C = -\rho\bar{o}_{B'}\varepsilon_{AC}, \tag{25}$$

where[19],[30] σ and $\rho(=\bar{\rho})$ measure, respectively, the shear and convergence of N, and are thus analogous to Σ and P in the case of H (eqn (22)). They are invariants of I-geometry. The derivatives

$$\text{þ}\sigma = 2\rho\sigma + \Psi, \qquad \text{þ}\rho = \rho^2 + \sigma\bar{\sigma} + \Phi \tag{26}$$

(Sachs equation for N[30], finite parts of curvature only being taken) define

$$\Psi = \Psi_{ABCD}\, o^A o^B o^C o^D, \quad \Phi = \Phi_{ABC'D'}o^A o^B \bar{o}^{C'}\bar{o}^{D'} \tag{27}$$

as invariants of II-geometry. The combination $\rho\Psi - \sigma\Phi$ is shown to be an invariant of I-geometry by þ-differentiation of the $(2,-2)$ quantity σ/ρ.† Furthermore, it is a consequence of eqn (40) that

$$\Psi_A = \Psi_{ABCD}o^B o^C o^D, \quad \Phi_{D'} = \Phi_{ABC'D'}o^A o^B \bar{o}^{C'} \tag{28}$$

are invariants of III-geometry. In passing, two equations relating various quantities invariant under III-geometry may be noted:

$$\eth_{A'B}\rho - \bar{\eth}_{BA'}\sigma = o_{B'}\Phi_{A'} - \bar{o}_{A'}\Psi_B, \tag{29}$$

$$o_B\text{þ}\Phi_{A'} - \bar{o}_{A'}\text{þ}\Psi_B - \eth_{A'B}\Phi + \bar{\eth}_{BA'}\Psi = 2\rho o_B\Phi_A - 4\rho\bar{o}_{A'}\Psi_B + 2\sigma\bar{o}_{A'}\Phi_B. \tag{30}$$

I should also remark that þ is not independent of $\eth_{AB'}$ since ($\eta^{\cdots}_{\cdots}$ having type (p,q))

$$\bar{o}^{B'}\eth_{B'A}\eta^{\cdots}_{\cdots} = o_A\text{þ}\eta^{\cdots}_{\cdots} - q\rho o_A\eta^{\cdots}_{\cdots}. \tag{31}$$

Note, also, that

$$o^A\eth_{B'A}\eta^{\cdots}_{\cdots} = -p\sigma\bar{o}_{B'}\eta^{\cdots}_{\cdots}. \tag{32}$$

---

† The types of the various quantities encountered here should be noted: o_A:(1,0); $\bar{o}_A$:(0,1); ρ:(1,1); σ:(3,−1); Φ:(2,2); Ψ:(4,0); $\Phi_{A'}$:(2,1); Ψ_A:(3,0); $\eth_{A'B}$:(2,0); þ:(1,1); $\phi_{AB'}$:(0,0); ψ_{AB}:(−1,1); ϕ:(−1,−1); ψ:(−3,1).

The various operators $\eth_{A'B}$, $\bar{\eth}_{AB'}$, þ do not commute. An example of a commutator (for scalar η) is

$$(\text{þ}\eth_{B'A}-\eth_{B'A}\text{þ})\eta=(\rho\eth_{B'A}+\sigma\bar{\eth}_{AB'}-p\bar{o}_{B'}\Psi_{A}-qo_{A}\Phi_{B'})\eta. \tag{33}$$

Two examples of invariant operations for III-geometry are

$$(\rho\eth_{E'C}-\sigma\bar{\eth}_{CE'})\xi_{A}+\xi_{G}o^{G}(\mathrm{X}_{ABCD}o^{B}o^{D}\bar{o}_{E'}-\Phi_{ABE'F'}o^{B}\bar{o}^{F'}o_{C}) \tag{34}$$

and, for ξ_A and $\eta_{A'}$ of the same type,

$$\eta_{E'}\bar{o}^{E'}\bar{o}_{D'}\eth_{C'A}\xi_{B}-\xi_{F}o^{F}o_{B}\bar{\eth}_{AC'}\eta_{D'}. \tag{35}$$

Finally, it should be remarked that, while the Gaussian curvature of any two-surface S of cross-section of N is obviously defined by the I-geometry of N, the complex Gaussian curvature,[26],[17] whose imaginary part defines an extrinsic imbedding invariant of S in N, is actually an invariant of III-geometry.

Let us now return to consideration of impulsive waves, but only in the cases for which the III-geometry induced by M^+ agrees with that induced by M^- (e.g. purely gravitational waves). The jump in the curvature tensor of M across N, can be obtained from the Bianchi identity

$$\nabla^{A}_{P'}\mathrm{X}_{ABCD}=\nabla^{Q'}_{B}\Phi_{CDQ'P'}. \tag{36}$$

Since this requires derivatives of δ-functions, it will be convenient to make some simplifications in the choice of l^a. Let v be a coordinate on M such that the hypersurfaces $v=$ constant are null, N being given by $v=0$. We choose the spinor field o^A so that

$$\nabla_{a}v=l_{a}=o_{A}\bar{o}_{A'}. \tag{37}$$

Then $\nabla_{[a}l_{b]}=0$, whence

$$0=\nabla_{A(B'}(\bar{o}_{C')}o^{A})=o^{A}\nabla_{A(B'}\bar{o}_{C')}+\bar{o}_{(C'}\nabla_{B')A}o^{A}. \tag{38}$$

We have $\Theta=\theta(v)$, $\Delta=\delta(u)$; then putting $\Delta'=\delta'(u)$ we get

$$\nabla_{AA'}\Delta=o_{A}\bar{o}_{A'}\Delta'. \tag{39}$$

Substituting the δ-function curvatures (21), together with the various terms involving Θ, into eqn (36) we see, using eqn (39), that the terms in Δ' go out; so after some simplifications entailed by eqns (38) and (23) we obtain the following expression for the jump in curvature across N:

$$[\mathrm{X}_{ABCD}o^{A}\bar{o}_{P'}-\Phi_{CDQ'P'}\bar{o}^{Q'}o_{B}]=2(\sigma\psi-\rho\phi)o_{(C}\varepsilon_{D)B}\bar{o}_{P'}+o_{C}o_{D}(\eth_{P'B}\psi-\bar{\eth}_{BP'}\phi). \tag{40}$$

Note that eqn (40) is written entirely in terms of invariantly weighted operations. It thus remains true if we drop eqns (37), (38), and (39) by rescaling o^A according to eqn (7).

For a purely gravitational wave, propagating through a vacuum, we can re-express eqn (40) as

$$[\Psi_{ABCD}o^{D}\bar{o}_{D'}] = o_{(A}o_{B}\eth_{C)D'}\psi. \tag{41}$$

This has a number of immediate implications. For example, if (and only if) $\eth_{A'B}\psi = 0$, the jump in the Weyl curvature is (Petrov) type {4} or {–}. This includes as a special case the well-known result[27] for gravitational shock waves ($\psi = 0$) that the curvature jump is type {4}. It also shows that for a purely gravitational impulsive wave, for which M^- and M^+ are flat, we must have $\eth_{D'C}\psi = 0$, whence (upon transvection with o^C, compare eqn (32)) we obtain $\sigma\psi = 0$. Thus, for a non-trivial, purely impulsive gravitational wave, the hypersurface N must be shear-free ($\sigma = 0$). But the only shear-free null hypersurfaces in flat space–time are actually portions of null hyperplanes or null cones. This is the basis of the statement made earlier, that the explicit examples considered, represented the only possible purely impulsive gravitational waves. These considerations also have the implication that, for example, if two purely impulsive gravitational plane waves collide, then they cannot remain purely impulsive when they separate, since each wave must pick up shear as it passes through the other.[31] (Thus, the waves acquire tails, i.e. produce back-scattering.)

Transvecting eqn (41) with o^C, we obtain (from eqn (32)) a formula

$$[\Psi_{ABCD}o^{C}o^{D}] = \sigma\psi o_{A}o_{B} \tag{42}$$

which (assuming $\psi \neq 0$) makes it clear that $\sigma = 0$ if and only if the jump in Ψ_{ABCD} has at least a triple principal null direction along l^a (type {31} or more special). In the general case, $[\Psi_{ABCD}]$ has a double principal null direction along l^a. Transvecting eqn (41) with $\bar{o}^{D'}$ and using eqn (31) we obtain the propagation equation

$$\text{þ}\psi = \rho\psi \tag{43}$$

for any impulsive gravitational wave.†

It is worth making a comparison with Maxwell theory here. Writing $F_{ab} = \gamma_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\bar{\gamma}_{A'B'}$ where γ_{AB} is symmetric, we obtain Maxwell's source-free equations in the form $\nabla^{AP'}\gamma_{AB} = 0$. Suppose the field is an electromagnetic purely impulsive wave. Then

$$\gamma_{AB} = \eta_{AB}\Delta \tag{44}$$

for some smooth symmetric η_{AB}, where Δ is defined, as before, with respect to a null hypersurface N. Let the background space–time be chosen to be smooth. (Thus we are considering a test electromagnetic field. Otherwise we should have difficulties, since the Maxwell stress tensor involves a square of a δ-function.) Substituting eqn (44) into Maxwell's equations, we get a term in Δ', namely $\eta_{AB}o^A\bar{o}^{P'}\Delta'$, whence $\eta_{AB}o^A = 0$ on N. Thus, the field is *null*: $\eta_{AB} = \eta o_A o_B$, so we can now write

---

† Provided $\sigma = 0$, the remaining condition on ψ implied by $\eth_{A'B}\psi = 0$ states that the complex quantity ψ is a holomorphic function of an appropriate complex coordinate ζ defined on cross-sections of N.

$$\gamma_{AB} = \eta o_A o_B \Delta. \tag{45}$$

When eqn (45) is substituted into Maxwell's equations we get no term in Δ' but a term $\Delta\nabla^{AP'}(\eta o_A o_B)$, showing that $\nabla^{AP'}(\eta o_A o_B) = 0$ on N. Simplifying the expansion of this, using eqns (38) and (23), we obtain

$$\eth_{A'B}\eta = 0; \tag{46}$$

two consequences of which are

$$\sigma\eta = 0, \qquad \text{þ}\eta = \rho\eta \tag{47}$$

(transvection with o^B and $\bar{o}^{A'}$ respectively), exactly as in the gravitational case. (In fact, if γ_A were a neutrino field satisfying Weyl's equation $\nabla^{AP'}\gamma_A = 0$, we should be led to $\gamma_A = \eta o_A \Delta$ with η satisfying equation (46) as in the electromagnetic and gravitational cases. However here $\sigma \neq 0$ can occur.)

From eqn (47), we see that for a purely impulsive electromagnetic wave (i.e. Huyghens's principle holding) we must have $\sigma = 0$. If this is to be possible for all light cones N in the space M, we deduce from eqn (26) that $\Psi^{\cdot} = 0$ for every point and for every choice of o^A. Hence the space must be conformally flat.

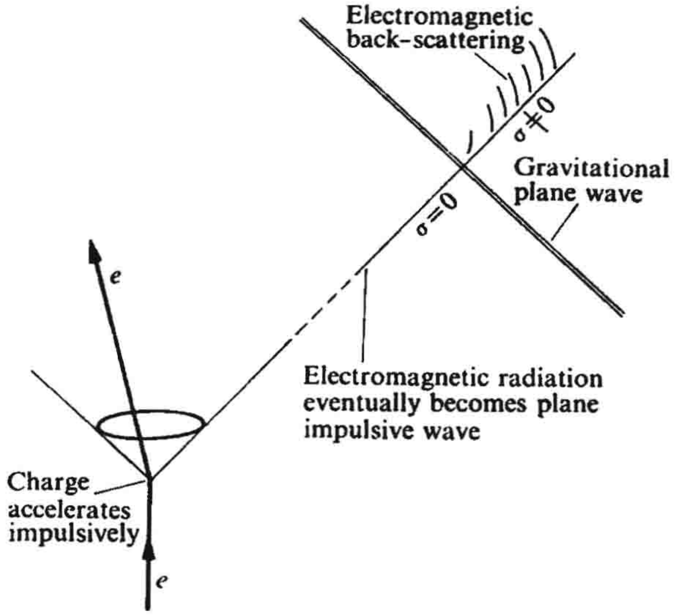

Electromagnetic back-scattering
$\sigma \neq 0$
Gravitational plane wave
$\sigma = 0$
e
Electromagnetic radiation eventually becomes plane impulsive wave
Charge accelerates impulsively
e

FIG. 4. The charge e receives an impulsive acceleration. The resulting retarded radiation approximates a purely impulsive plane electromagnetic wave in the far zone, but after an encounter with a gravitational plane wave, the electromagnetic wave back-scatters and so ceases to satisfy Huyghens's principle. If a scalar field were to replace the electromagnetic field, then back-scattering would not occur in this example.

This is an affirmative form of Hadamard's conjecture (i.e. that Huyghens's principle implies conformal flatness).[32] The corresponding result for a scalar wave field (subject to the conformally invariant wave equation $(\nabla_a\nabla^a+4\Lambda)\alpha = 0$) is known to be false when M is a gravitational plane wave.[12],[14] That is, the field α can be a pure δ-function on any light cone in a plane-wave space–time M, even though the cones pick up shear ($\sigma \neq 0$).

The distinction in the case of a Maxwell field can be illustrated by the situation indicated in Fig. 4. A charge e in a flat portion of the space–time receives an impulsive acceleration. It radiates retarded electromagnetic field which, if received far enough away, becomes effectively a pure impulsive electromagnetic wave, provided it remains in flat space–time. But if the wave finally crosses, say, a gravitational plane wave, then the wave front picks up shear ($\sigma \neq 0$) so that the wave back-scatters and can no longer be clean cut (i.e. is not purely impulsive).†

Finally, it may be pointed out that if the above wave were a scalar (test) field α and if the gravitational plane wave were of 'sandwich' form,[35] then the clean-cut nature of the emergent scalar wave would be closely related to a phenomenon discovered by Friedlander,[36] according to which purely impulsive scalar waves in flat space–time can reside on certain shearing null hypersurfaces constructed from Dupin cyclides. In this case, N becomes such a hypersurface.

REFERENCES

1. LANCZOS, C. *Phys. Z.* **21,** 540 (1922).
2. O'BRIEN, S. and SYNGE, J. L. *Communs Dubl. Inst. advd Stud.* A **9** (1959).
3. STELLMACHER, K. *Math Annln* **115,** 740 (1938).
4. SYNGE, J. L. *Relativity: the general theory.* North-Holland, Amsterdam (1960).
5. TAUB, A. H. *Illinois J. Math.* **1,** 370 (1957).
6. LICHNEROWICZ, A. *Théories relativistes de la gravitation et de l'électromagnetisme.* Masson, Paris (1955).
7. PAPAPETROU, A. and TREDER, H. J. *Math. Nachr.* **20,** 53 (1959).
8. TREDER, H. J. *Gravitative Stosswellen.* Akademie-Verlag, Berlin (1962).
9. DAUTCOURT, G. *Math. Nachr.* **27,** 277 (1964).
10. DAUTCOURT, G. *Math. Nachr.* **42,** 309 (1969).
11. ISRAEL, W. *Nuovo Cim.* B **44,** 1 (1966).
12. GÜNTHER P. *Arch. ration. Mech. Analysis* **18,** 103 (1965).
13. GÜNTHER, P. *Wiss. Z. Karl-Marx-Univ. Lpz.* **14,** 497 (1965).
14. MCLENAGHAN, R. G. *Proc. Camb. phil. Soc. math. phys. Sci.* **65,** 139 (1969).
15. DAUTCOURT, G. *J. math. Phys.* **8,** 1492 (1967).
16. DAUTCOURT, G. *Math. Nachr.* **36,** 311 (1968).

---

† There is an apparent conflict here with the work of Künzle.[33],[34] This may be due to a difference in interpretation of Huyghens's principle.[13]

17. Penrose, R. Null hypersurface initial data for classical fields of arbitrary spin and for general relativity. In *Quantization of generally covariant fields* (A.R.L. Technical Documents, Report no. 63–56, ed. P. G. Bergmann). (1963).
18. Sachs, R. K. *J. math. Phys.* **3,** 908 (1962).
19. Penrose, R. Structure of space–time. In *Battelle rencontres* (ed. C. M. De Witt and J. A. Wheeler). Benjamin, New York (1968).
20. Penrose, R. *Int. J. theor. Phys.* **1,** 61 (1968).
21. Brinkmann, H. W. *Math. Annln* **94,** 119 (1925).
22. Ehlers, J. and Kundt, W. Exact solutions of the field equations. In *Gravitation* (ed. L. Witten). Wiley, New York (1962).
23. Robinson, I. Report to Eddington group. Cambridge (1956).
24. Rosen, N. *Phys. Z. SowjUn.* **12,** 360 (1937).
25. Robinson, I. and Trautman, A. *Proc. R. Soc.* A **265,** 463 (1962).
26. Geroch, R. P., Held, A., and Penrose, R. A space–time calculus based on pairs of null directions. Preprint (1970).
27. Pirani, F. A. E. *Phys. Rev.* **105,** 1089 (1957).
28. Penrose, R. Gravitational energy flux and elementary optics. In *Perspectives in geometry and relativity* (ed. B. Hoffmann). Indiana University Press, Bloomington (1966).
29. Newman, E. T. and Penrose, R. *J. math. Phys.* **7,** 863 (1966).
30. Sachs, R. K. *Proc. R. Soc.* A **264,** 309 (1961).
31. Khan, K. A. and Penrose, R. *Nature, Lond.* **229,** 185 (1971).
32. Hadamard, J. *Lectures on Cauchy's problem in linear partial differential equations*. Dover, New York (1952).
33. Künzle, H. P. The influence of space–time curvature on the electromagnetic field of a charged particle. Ph.D. Thesis, King's College, London (1967).
34. Künzle, H. P. *Proc. Camb. phil. Soc. math. phys. Sci.* **64,** 779 (1968).
35. Bondi, H., Pirani, F. A. E., and Robinson, I. *Proc. R. Soc.* A **251,** 519 (1959).
36. Friedlander, F. G. *Proc. Camb. phil. Soc. math. phys. Sci.* **43,** 360 (1947).

56

ON THE NATURE OF QUANTUM GEOMETRY

Magic without magic (J. A. Wheeler Festschrift), pp. 333–54, Freeman, 1973

Spin networks (see Chapter 31) appear again, here discussed in relation to the symplectic struture of twistor space (Chapters 37 and 52).

On the nature of quantum geometry. *Magic without magic* (J. A. Wheeler Festschrift), pp. 333–54, Freeman, 1973

ON THE NATURE OF QUANTUM GEOMETRY

by Roger Penrose
Dept. of Mathematics, Birkbeck College London

As a way of honouring Professor Wheeler's 60th birthday, I propose to take this opportunity of elaborating upon certain somewhat speculative ideas which I have tried to hint at on occasion, concerning the possible nature of a quantized space-time. The reader will not need to be too discerning to recognize some substantial differences between the ideas I am proposing here and those which Professor Wheeler has, on many occasions, so eloquently and forcefully put forward. Nonetheless, there is little doubt in my own mind as to the very great inspirational influence that Professor Wheeler's own views have had in the development of several of the thoughts which I am attempting to set forth here.

To begin with, let me make clear that I do not necessarily mean by "quantized space-time" something which could be obtained by applying standard (or even non-standard) techniques of quantization to Einstein's general theory of relativity. What I wish to say has its roots in something which is really more primitive than either quantum theory or relativity as such. This is the question of the fundamental role played by the mathematical concept of continuum in virtually the whole of accepted present-day physical theory. Not only does the continuum occupy a basic position in our mathematical models of *space* and *time* (with the concomitant implication of a continuous nature for many related physical concepts such as velocity, energy, momentum, temperature, etc.), but so also does present-day *quantum theory* rest crucially on a continuum concept, namely on the two-dimensional complex continuum of probability amplitudes, this being the continuum which also occurs in the superposition law.

Let me say at the outset that I am not happy with this state of affairs in physical theory. The mathematical continuum has always seemed to me to contain many features which are really very foreign to physics. This point has been argued forcefully, particularly by Schrödinger[40] and also by a number of other physicists and philosophers.[29] If one is to accept the physical reality of the continuum, then one must accept that there are as many points in a volume of diameter 10^{-13} cm, or 10^{-33} cm, or 10^{-1000} cm, as there are in the entire universe. Indeed, one must accept the existence of more points than there are natural numbers, between any two points in space no matter how close together they may be. (And we have seen that quantum theory cannot really eliminate this problem, since it brings in its own complex continuum.) It seems clear that such "points" have actually very little to do with physical reality. Nevertheless their postulated existence is virtually essential to contemporary physical theory. The concepts of open and closed sets, for example, depend vitally on this continuum of points and are essentially meaningless in strictly physical terms; but such concepts can be used to great effect when mathematical *theory* is applied to physics (e.g. in proving global "singularity" theorems about space-times etc.).[19,20,33]

I think it must be the case that the all-pervading use of the continuum in physics stems from its *mathematical utility* rather than from any essential physical reality that it may possess. However, it is not even quite clear that such use of the continuum is not, to some extent a *historical accident*. For although the essential mathematical ideas can be traced back to Eudoxus (4th century B. C.), it was not until sometime after Newton and Leibniz invented the calculus that it was felt to be necessary to formalize and make completely rigourous the mathematical continuum concept. But there are other "non-standard" continua,[37,39] different from the one that has now become conventional, which could equally well have been adopted in order to make the calculus rigourous. In these non-standard continua, "infinitesimal" and "infinite" elements are introduced and are treated as being just as "real" as the "real numbers" of conventional analysis. In fact, it has occasionally been argued that non-standard analysis might really have been a more natural development of the ideas of Newton, Leibniz and Euler than the actual analysis which Cauchy, Weierstrass and others finally formalized. (In addition, non-standard analysis allows one to define such concepts as Dirac delta-functions so that they become effectively "ordinary" functions.[25]) If the history of mathematics had developed differently, then we might, by now, have formed a very different view, from the one now prevalent, of the nature of space and time, and of many other physical concepts.

I do not want to imply here that non-standard analysis ought to be employed in physical theory. I wish merely to point out the lack of firm foundation for assigning any physical reality to the *conventional* continuum concept. My own view is that ultimately physical laws should find their most natural expression in terms of essentially *combinatorial* principles, that is to say, in terms of finite processes such as counting or other basically simple manipulative procedures. Thus, in accordance with such a view, should emerge some form of discrete on combinatorial space-time. I do not mean that necessarily we should arrive at a space-time containing a discrete set of points, such as the lattice space-time of Schild,[38] or some discrete causal space,[24] or, for example, some structure based on Ahmavaara's large finite field.[2] I would expect, rather, that the concept of a space-time composed of *points* should *cease* to be an appropriate one - except in some kind of limiting sense. But if points are not to be the basic elements of the discrete space-time, then how are we to decide what these basic elements should in fact be?

It is my view that an essential insight into the nature of the appropriate combinatorial structure should actually emerge once the interrelations between quantum physics with the present-day view of space-time are fully appreciated.* But was I not arguing that quantum theory is of no value for eliminating the continuum, since it entails the use of the complex continuum right at the outset? This is certainly true as it stands. However, I think it is important to make the distinction, here, between quantum theory (which requires a complex continuum concept) and quantum physics (according to which certain physical quantities are recognized as being actually discrete, while having previously been taken to be continuous). My idea is to try to "reformulate" physical laws so that they may be expressed entirely in terms of quantities which are discrete according to quantum physics. These "reformulated" laws would, hopefully, be expressible entirely in combinatorial terms, even though they would be essentially re-expressing the

---

[*] I should remark, here, on the interesting programme due to D. Finkelstein[4]; according to which quantum theory is to be built into the mathematical description of space-time even at basic level of set theory. To be in accordance with the viewpoint I am expressing here, however, such a "quantization of set theory" would have to be accompanied by a suitable "combinatorialization of quantum theory".

content of conventional quantum theory, of space-time theory and, perhaps, of other aspects of physics as well. Thus the quantum theory and space-time theory would be expected to arise together, out of some more primitive combinatorial theory.

In order to be more explicit as to the sort of "reformulation" that I have in mind, I should describe a certain model which I have referred to elsewhere[32,35] and which may be thought of as a prototype for this type of theory. The basic idea of the model is to take the concept of total angular momentum and to regard this as the primary physical quantity; then the quantum mechanical rules for combining (non-relativistic) total angular momenta can be re-expressed in purely combinatorial terms; finally the concept of space direction is to be extracted and shown to agree, in the limit of large angular momenta, with the ordinary geometry of directions in a Euclidean three-dimensional space. The reason for fastening attention on the concept of total angular momentum, in the first instance, is that it seems to be more or less uniquely singled out by a number of criterea. For we have to choose something, which according to quantum physics, is discrete, preferably taking numerical values which are integer multiples of some basic unit (in this case $\frac{1}{2}\hbar$). Since we are interested in reconstructing a form of space out of this discrete physical quantity, we need something which is intimately related to spatial and directional properties. This suggests that we should use angular momentum, rather than other possible quantum numbers. Finally, since we wish to construct space rather than depend upon any assumption of pre-existing spatial directions, we must use total angular momentum (j-values) rather than the angular momentum in some preassigned direction (m-values).

The model operates with structures I call spin-networks. An example is illustrated in figure 1. The line segments are called units and are to be loosely interpreted as "world lines" of particles, or of simple systems which may be regarded as momentarily isolated from the rest of the universe. Each of these particles or simple systems has to be stationary (not really moving relative to the others, in this model) and possess a well-defined total angular momentum. The integer labelling each unit is its spin-number. This measures the total angular momentum of the unit, as a multiple of $\frac{1}{2}\hbar$. Exactly three units must come together at each (internal)

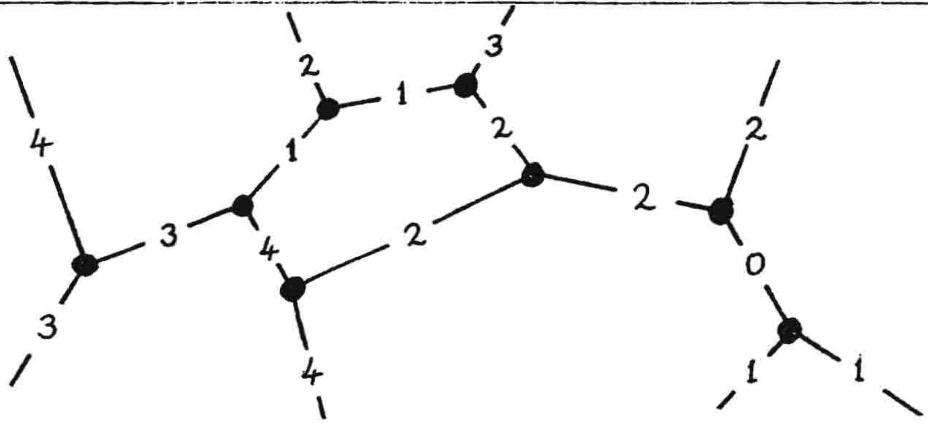


FIGURE 1
An example of a spin-network. The numerals denote total spin values in units of $\frac{1}{2}\hbar$.

vertex. Depending upon which way we choose to regard time as progressing in the diagram (but normally we choose upwards, conventionally) we may interpret the meanings of the vertices in different ways. We may think of a vertex as representing the combining of two units together to make a third, or as the splitting of a single unit into two separate units. A unit which is not terminated at both ends by a vertex is called an end-unit.

Given a spin-network we may calculate, using purely combinatorial means, a certain non-negative integer, called the norm of the spin network. I have described this calculational procedure elsewhere[35], and I do not propose to go into the details here. Suffice it to say that this combinatorially defined norm may be used to obtain the probability that the spin number x is the result, when two end-units of a given spin-network α, come together to form a new unit.

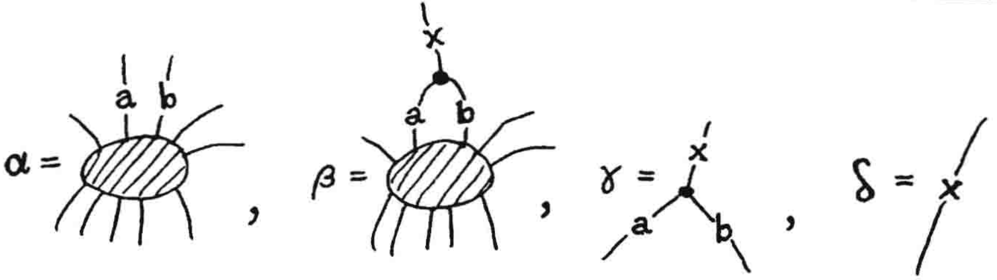


FIGURE 2
The a- and the b-units come together to form an x-unit. With spin-networks labelled as above, the probability of the value x is given by Eq(1).

With spin-networks labelled as in figure 2, the formula turns out to be:

$$\text{probability of } x = \frac{\text{norm } \beta \text{ norm } \delta}{\text{norm } \alpha \text{ norm } \gamma}$$

$$= \frac{\text{norm}\,\beta\,(x+1)\Big\{\{\tfrac{1}{2}(a+b+x)+1\}!}{\text{norm}\,\alpha\Big\{\{\tfrac{1}{2}(a+b-x)\}!\,\{\tfrac{1}{2}(b+x-a)\}!\,\{\tfrac{1}{2}(x+a-b)\}!\Big\}^{-1}} \qquad (1)$$

which is a rational number. This result is obtained from conventional non-relativistic quantum mechanics, but things have to be reformulated considerably in order to be able to state the result in a reasonably simple purely combinatorial form. I should mention, also that a spin-network represents a physical process which is forbidden (i.e. zero probability) by the rules of non-relativistic quantum mechanics if and only if its norm vanishes.

The proposal now is to extract a concept of space from this scheme - or, rather, the simpler concept of directions in space, from this particular model. The basic idea stems from the quantum mechanical fact that a system with zero total angular momentum must by spherically symmetrical, and so cannot be used to define a direction in space; a system of total spin $\frac{1}{2}\hbar$ is not much better, since it "sees" but two alternative "directions" available to it as regards its state of spin; for spin $\hbar$ there are but three alternative "directions" and so on. Only for a system involving a comparatively large total spin value, can we expect that it could define a direction in any well-defined way. And once we have a large spin, we can envisage a fairly well-defined rotation axis as a convenient means of defining us a direction in space.

Next, angles between rotation axes can be defined in terms of certain simple "experiments". A spin $\frac{1}{2}\hbar$ unit is detached from some unit of large spin (called a large unit) and then reattached to another large unit. Let us suppose that the spin of the first large unit is reduced by $\frac{1}{2}\hbar$. Then the spin of the second large may be increased by $\frac{1}{2}\hbar$ — with probability p, say — or it may be reduced by $\frac{1}{2}\hbar$ — with probability $1 - p$. If there is a well-defined angle θ between the spin axes of the two units, then in accordance with standard quantum mechanics we have (in the limit of large spin)

$$p = \cos^2 \tfrac{1}{2}\theta \qquad (2)$$

We can use (2) as the definition of the angle between two large units, provided that the two units have, in some appropriate sense, a well-defined angle between them. (Without such a proviso this definition would not lead to a reasonable geometry.)

Before considering the question as to when the angle between the spins of two large units can be considered as "well-defined". I should first be a little more precise as to the interpretations of the probabilities arising here. These probabilities are always calculated starting from a given spin-network α (cf. figure 2) which is supposed to represent some known portion of the universe. The spin-network α may occur again at various other places in the universe. At some places it may be part of a more extended spin-network β ; at other places it may be part of β' (which is to differ from β only in that one of the spin-numbers is different: cf. x in figure 2) ; at other places it may be part of other spin-networks. The calculated probability is then supposed to give us the relative frequency of β to β' (given α) in the universe[*]. Now, it may be that α contains two large end-units, M and N , but without much connecting network between them.

---

[*] This point of view serves to make the probability concept tolerably precise, for purposes of the model. It may possibly leave something to be desired for a physically more realistic model. For example, whether or not the units labelled a and b in figure 2 could actually come together to form another unit might depend on whether or not a suitable "particle" with spin-number x really existed. This would influence the probabilities in a way not taken into consideration here.

This is the situation where we "know" rather little about the relationship between M and N , so the probability p that we calculate in the above experiment cannot really be thought of as defining an "angle" according to (2) , but is rather partly a measure of our ignorance of the relation between M and N . For the angle between M and N to be well-defined, we should require this ignorance factor to be very small.

But how are we to decide whether a probability arises partly from ignorance or entirely from angle? Consider the above experiment repeated in the way depicted in figure 3. The second experiment gives a probability p'

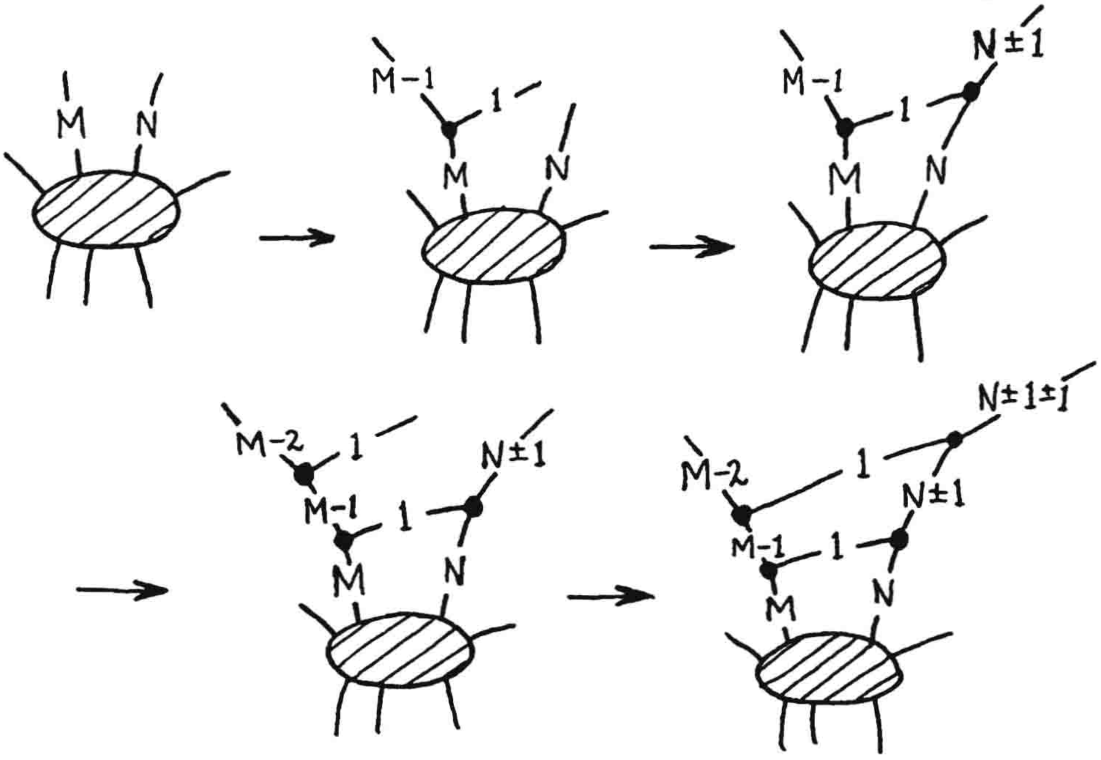


FIGURE 3

An angle measuring experiment, repeated in order to eliminate the "ignorance factor".

which could be affected by the result of the first experiment. If p is *not* significantly affected by the result of the first experiment (and $p \hat{=} p$) then we can say that the ignorance factor is small and the angle between M and N is well-defined. (If the angle between M and N is not well-defined in this sense, then we can generally make the angle better-defined by carrying out a fairly large number of similar experiments in succession. The probability normally settles down to some fixed value.) Now, the following theorem can be proved: *if a spin-network has a number of large end-units such that the angle between any two of them is well-defined in the above sense, then these angles can be consistently interpreted as angles between directions in a Euclidean three-dimensional space.*

This result is very satisfactory, but in a way it is perhaps *too* satisfactory! It involves a peculiar feature which I want to emphasize particularly in this article. Suppose we had set up the normal quantum mechanical formalism for the description of the situation, that is, by giving states in terms of wave functions involving, say, coordinates r, θ and ϕ, these being ordinary spherical polar coordinates for a preassigned background Euclidean space. Can we identify the directions (of spin-axes) that we end up with *as* the directions in this background Euclidean space? No, we certainly cannot! The condition, given above, that angles between spin-axes of units be well-defined, by no means ensures that the spin-axes themselves correspond to well-defined directions

(For example, a state with $m = 0$ does not define a good background direction for its spin-axis, no matter how large j is.) I propose to take the attitude, however, that in this model it is the geometry of "spin-axes" of the large units which is the *real* geometry. The background space, with its spherical polar r, θ and ϕ has no actual physical meaning and is introduced merely as a convenience for calculations (that is, if one chooses to use the conventional wave-function description, rather than a combinatorial procedure). Thus, *the system itself defines the geometry* and the background space is really an irrelevance.

It appears that we may think of the relation between the background space and the real space as being given by a *unitary transformation in Hilbert space*. Thus, states which give well-defined directions in one space might correspond to linear superpositions of states having well-defined directions in the other. This idea is very close to a suggestion due to Aharonov[1], according to which one can pass from one concept of geometry to another by applying a Hilbert space unitary transformation. On this view, an electron moving through two slits, for example, does not "feel itself to be split" since the geometry "felt" by the electron is not quite the same as that of the slits. According to the electron's geometry (which would be related to the background geometry by a unitary transformation) the electron would remain intact while the background space would be "split" in a certain sense.

Also it is not actually necessary that the "real" geometry should fit together globally to give a space of the same kind as the background space. Let me illustrate what I have in mind by reference to the model. Consider a spin-network α which consists of two portions λ and μ, each of which has a number of large end-units with a well-defined angle between them (figure 4). Thus, each of λ and μ defines its own Euclidean geometry of directions in a well-defined way. Now suppose that the connections between λ and μ in α are not sufficient (or not of the right type) to ensure that the geometry defined by λ and that defined by μ are consistent with one another (i.e. the angle between a large unit of λ and a large unit of μ need not be well-defined). Then the large units of λ need not correspond to states of spin with wlll-defined rotation axes according to the μ geometry and *vice versa*. Thus, we may expect that the relation

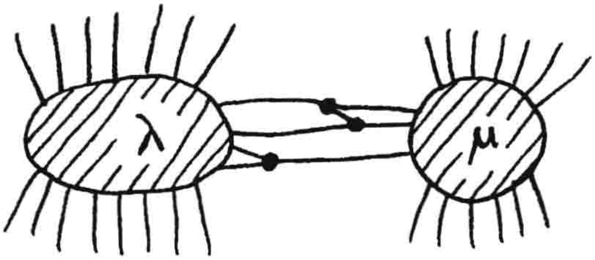


FIGURE 4

The spin-networks λ and μ each define their own geometry, but owing to inadequate connections between λ and μ, the two geometries are not consistent with one another.

between the λ-geometry and the μ-geometry should occur via something like a unitary transformation in Hilbert space.

The circumstance whereby each part of a structure can define its own local Euclidean geometry, but where these local geometries need not quite patch together to make a global Euclidean geometry, is reminiscent of the situation occurring with Riemannian geometry, as used in general relativity. In fact, it is one of my main contentions in this article that the curvature of space-time may indeed have its origins in an effect of the kind just described. But how are we to relate, in any meaningful way, the effects of space-time curvature to the effects of a unitary Hilbert space transformation? On a classical level, we would expect such unitary transformations to show up as canonical transformations in suitable classical variables. Thus, it must be possible to relate space-time curvature to canonical transformations in a suitable manner. I shall show how this can be done, first, in a way which is closely related to (but not quite the same as) the way which I believe to have the most significance. I shall describe the way I prefer afterwards.

Let M be a space-time manifold* and let C be its cotangent bundle.[23,27] Thus, each point of C represents a point x of M together with some covariant vector p_a (referred to as a momentum vector) at the point x . The eight-dimensional manifold C has a natural symplectic (or "Hamiltonian") structure defined by the two-form[15,27]

---

*A four-dimensional, pseudo-Riemannian,(+---), time-oriented Hausdorff manifold with a C^2-metric.

$$\omega = dp_a \wedge dx^a \tag{3}$$

Of interest also is the naturally defined one-form

$$\phi = p_a \, dx^a \tag{4}$$

of which ω is the exterior derivative: $\omega = d\phi$. The inverse of the tensor (on C) defining ω is the bivector which defines the Poisson bracket :

$$\{\ ,\ \} = \frac{\partial}{\partial p_i} \otimes \frac{\partial}{\partial x^i} - \frac{\partial}{\partial x^i} \otimes \frac{\partial}{\partial p_i} \ . \tag{5}$$

Any smooth transformation of C to itself which preserves the symplectic structure ω -- or equivalently which preserves the Poisson bracket (5) -- will be a canonical transformation. However, these are not quite the transformations that concern us here. We must first construct a reduced phase space by a standard procedure.[22] Let C_K be the seven-dimensional submanifold of C for which the (squared) Hamiltonian

$$H \equiv g^{ab} p_a p_b \tag{6}$$

takes the constant value $H = K$. The Poisson bracket now defines a vector field χ (or "flow") on C by

$$\chi(f) = \{H,f\} \tag{7}$$

(f being an arbitrary smooth function on C). The integral curves of χ are the "lifts" of geodesics

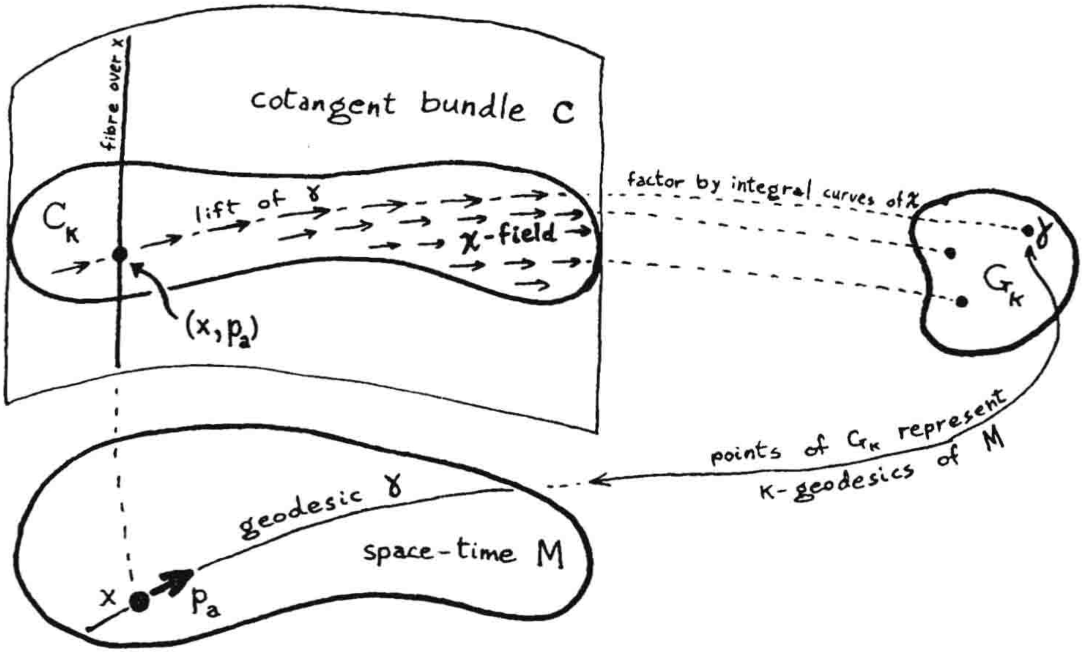

fibre over x
cotangent bundle C
C_K
lift of γ
χ-field
(x,p_a)
factor by integral curves of χ
γ
G_K
points of G_K represent K-geodesics of M
geodesic γ
space-time M
x
p_a

FIGURE 5

The space G_K of K-geodesics in M inherits a symplectic structure from that of the cotangent bundle C .

in the space-time M (cf. fig. 5), the tangent vector to the geodesics being $p^a = g^{ab}p_b$.[9,22] Now, the space C_K , being odd-dimensional, does not possess a symplectic structure. (The form ω induces a degenerate two-form called a pre-symplectic structure on C_K). However we can (locally, at least) factor out C_K by the integral curves of χ to obtain a symplectic six-dimensional manifold G_K. Each point of G_K represents a geodesic γ in M whose (parallelly propagated) tangent vector p^a has squared length $p^a p_a = K$. Call such a geodesic a K-geodesic.* The symplectic structure on G_K is that induced by the two-form ω.

It is worthwhile to examine the geometric meaning[9,34] of this symplectic structure on G_K, and also the meaning of ϕ in relation to G_K. In fact both ϕ and ω represent integrals of the Jacobi equation for geodesics. That is to say they represent properties of neighbouring geodesics which can be calculated at any one point but which are actually constant along the geodesics. Let us consider ϕ, first, as given by (4). We may think of p^a as the tangent vector to some K-geodesic γ and dx^a as a "connecting vector" which connects a point x on γ to a corresponding neighbouring** point x' on a neighbouring K- geodesic γ' to γ (see fig. 6). The fact that $p_a dx^a$ is constant* along γ is a well-known property of Lie derivatives.[23] Now consider the interpretation of ω In "old-fashioned" notation we write

---

* We can, of course, restrict attention to the three values $K = 1,0,-1$ if desired, since all other cases are related to these by scalings.

** For the purposes of such geometrical descriptions it is convenient to adopt an "old-fashioned" attitude to "dx^a" and to talk about "neighbouring" points and curves. There is actually no real conflict between the "old-fashioned" and "contemporary" viewpoints here. An equation such as $Q_a dx^a = 0$ can be interpreted either as "the infinitesimal vector dx^a connecting two neighbouring points x and x' is orthogonal to Q_a" or as "the one-form $Q_a dx^a$ maps to zero the vector at x with which we are concerned".

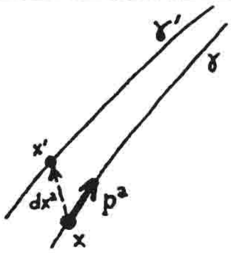


FIGURE 6

The quantity $\phi = p_a dx^a$ is constant along γ (where if $K \neq 0$, x' has the same parameter value on γ' as x has on γ).

$$\omega = dp_a \delta x^a - \delta p_a dx^a \tag{8}$$

where dx^a is as before (but where parameter values need not now correspond) and where δx^a is viewed as a "connecting vector" which connects the point x on γ to a neighbouring point x'' on a second neighbouring geodesic γ''. The tangent vectors to γ' and γ'' are to be, respectively, $p^a + dp^a$ and $p^a + \delta p^a$. (See figure 7.) The fact that (8) is constant along γ is sometimes referred to as the Lagrange identity[23]. It is a consequence of the interchange symmetry of the Riemann tensor: $R_{abcd} = R_{cdab}$. We may think of ω as defining a measure of the rotation of three neighbouring geodesics about one another.

---

* If $K = 0$, this does not require x and x' to occur at corresponding parameter values on γ and γ'. Thus the one-form ϕ carries over to G_k if $K = 0$.

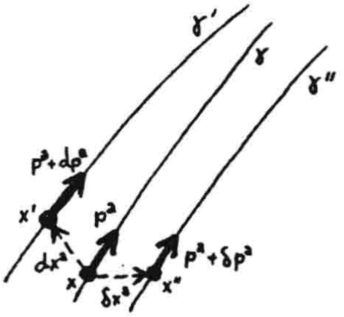


FIGURE 7

The geometric meaning of the symplectic structure of G_K : the rotation measure $\omega = dp_a \delta x^a - \delta p_a dx^a$ is constant along γ (Lagrange identity).

Let us return to the question of the relation between space-time curvature and canonical transformations. The canonical transformations are actually to be applied to the space of K-geodesics G_K (for some fixed K) , rather than to the space-time M . (Thus, we must expect that if a space is to be constructed out of some combinatorial principles in accordance with the general idea of the model described earlier, then this space should, in the first instance, be more like G_K than directly like M .) We can imagine that it should be possible to reconstruct the geometry of M once sufficient structure of the space G_K is known. For, we may associate each point x of M with the system $G_K(x)$ of K-geodesics through x . Each $G_K(x)$ may be viewed as a three-dimensional submanifold of G_K , these submanifolds forming a four-dimensional family within G_K . If we know this family of submanifolds, then we should be able to reconstruct the space-time M from G_K .

Let us consider how this can be done in the special case when M is Minkowski space-time. The flat geometry of M assigns considerably more structure to G_K than just the symplectic geometry of G_K. I do not want to go into much detail here, but the essential point is that the shear of a congruence of geodesics in M can, owing to the flatness of M, be defined in a way which refers to the geodesics in their entirity and does not depend on a choice of a particular point on the geodesic. It follows that the shear concept can be interpreted in terms of some local structure on G_K; in fact, in terms of some tensor field σ on G_K. One way of achieving this is by means of a convariant tensor on G_K of valence four, which, using a notation similar to that of (8) can be expressed as:

$$\sigma = dx^a dp_a \delta p^b \delta p_b - \delta x^a \delta p_a dp^b dp_b . \qquad (9)$$

In order to fix ideas, let us suppose that σ is in fact defined by equation (9), even though this is not completely satisfactory for later purposes. The form ω actually already (satisfactorily) achieves a corresponding interpretation with regard to the rotation of a congruence of geodesics (but, in this case, independently of the flatness).

The importance of being able to interpret both the shear and rotation on G_K is that σ and ω may now be used to locate the $G(x)$ submanifolds in G_K and hence to reconstruct space-time points. For if x is a point in M then the congruence of K-geodesics through x (now straight lines) has the property that its shear and rotation both vanish. (I am now supposing that $K \neq 0$, in order to keep the discussion simple.) Conversely, any congruence with vanishing shear and rotation (and nonvanishing divergence) is a $G_K(x)$ system and hence defines a unique point x in M.

But how does the effect of curvature in M show up in G_K? Let us first consider a very idealized situation. (This is not really essential, but it serves to clarify matters.) Suppose that M contains two (geodesically convex) regions F_1 and F_2 of flat space-time and suppose that some open set E in G_K represents K-geodesics passing through the interiors of both F_1 and F_2. The flat geometry of F_1 assigns some "shear structure" σ_1 to E; similarly the flat geometry of F_2 assigns some "shear structure" σ_2 to E. But, in general, we shall have $\sigma_1 \neq \sigma_2$. This is because a congruence of geodesics, shear-free in F_1, will normally begin to pick up shear as soon as it leaves F_1 and enters a region of curved space-time. Only in very exceptional circumstances would all the shear exactly cancel out by the time the geodesics finally enter the region F_2.

Now the K-geodesics through x , where x is some point in the interior of F_1 , will have the property that while within F_1 , they constitute a shear-free, rotation-free congruence. This fact could be recognized in E , by reference to σ_1 and ω . But there will be many other shear-free, rotation-free congruences of K-geodesic segments in F_1 . Some of these will appear to be converging on points which lie just *outside* the F_1 region. (See figure 8.) As soon as the geodesics enter the curved region they will begin to pick up shear, so that they will not normally converge cleanly on any actual point of M . Correspondingly, a congruence of K-geodesics which actually *does* converge on a point y just outside F_1 will normally possess a certain amount of shear while *in* F_1 .

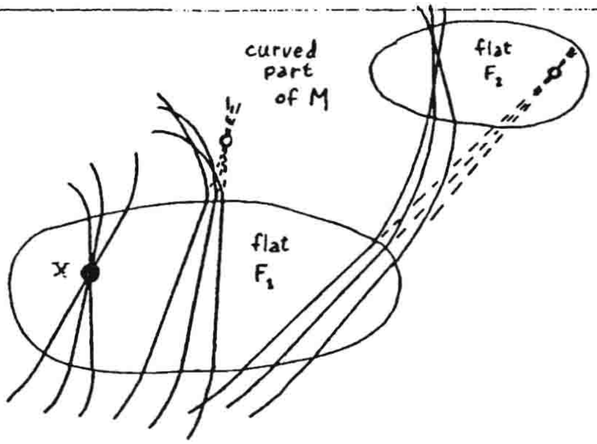


Figure 8

Points in F_1 may be located in terms of shear-free rotation-free congruences of geodesics, but on this basis, points outside F_1 would appear to be "fuzzed out".

Thus, we see that the "shear structure" σ_1 of E can be used to help locate the points of F_1 , but that if we are not careful we will also "locate" things which appear to be points outside F_1 , but which are actually not points of M at all. Similarly, σ_2 helps to locate the points of F_2 , but points which are not in F_2 are incorrectly "located" by this means. In particular, σ_1 locates points in F_1 correctly and points in F_2 incorrectly, while with σ_2 the situation is just the reverse. In a sense, the points of F_2 appear to be "fuzzed out" when viewed from F_1 (i.e. using σ_1) and vice versa. This "fuzzing out" may be validly thought of as being the result of a canonical transformation of E . Such a canonical transformation would preserve ω (the symplectic structure must be preserved by definition of "canonical") and could transform σ_1 to σ_2 . To achieve such a transformation explicitly, we can set up ordinary Minkowski coordinates in F_1 . We can then set up a related canonical coordinate system in a standard way, for the relevant part of the cotangent bundle C , and thence arrive at canonical coordinates for $E \subset G_K$. Similarly, ordinary Minkowski coordinates in F_2 will give rise to different canonical coordinates for E . These two coordinate systems for E will be related by a canonical transformation, the components of σ_1 with respect to the first being the same as the components of σ_2 with respect to the second.

Let us pass, now, to the case of a general curved space-time M . For simplicity suppose that $K \geq 0$ (so that the geodesics are all timelike or all null) and also that M is globally hyperbolic.[8, 26] Then M admits slicings by certain spacelike hypersurfaces, with the property that any one of them could be used globally as a Cauchy hypersurface for M[18]. Each S hypersurface will then intersect each K-geodesic once and once only. Thus, given S, we can examine the extent of shearing of a congruence of K-geodesics at the intersections of these geodesics with S. This gives us a definition of "shear structure" σ_S for G_K relative to the hypersurface S. If we wish to use (9) for the definition of σ in flat space-time, then we can still use (9) in curved space-time to define σ_S, by choosing dx^a and δx^a to connect neighbouring points lying within the hypersurface S (so dx^a and δx^a are tangent to S).

The six-dimensional manifold G_K will now possess the following structure. In the first place it will have a permanent symplectic structure defined by the two-form ω (and, if $K=0$, the one-form ϕ with $\omega = d\phi$). In the second place it will have a shifting "shear structure" σ_S, which depends on the location of the hypersurface S in M. If S is moved over a region of flat space-time in M, then σ_S will not change. If S is moved over curved regions in M, on the other hand, then σ_S will shift, its rate of change being actually governed by the Riemann tensor in M at the points over which S moves. It is σ_S (in relation to ω) which carries all the information as to the actual geometry of M. The form ω by itself conveys no information as to the metric structure of the space-time. Given σ_S , the points of S can be realized in G_K as three-dimensional submanifolds $G_K(x)$ whose tangent vectors annihilate σ_S and ω (the shear-free and rotation-free conditions). But the points of M which lie off S will, due to the curvature in M , generally appear to be "fuzzed out" from the point of view of the σ_S structure, in the sense that the $G_K(x)$ cannot now be recognized as shear-free, with respect to σ_S . The further away the points are from S , the greater, in general, will be this "fuzzing out". Since ω does not shift on G_K , we can think of this "fuzzing out" as a canonical transformation effect.

I want, now, to modify this picture somewhat, so as to bring it more into line with some other features which I feel should be involved, in connection with a quantized space-time. As a first step let me specialize to the case $K = 0$. I have several reasons for wishing to do this. One of these is that it turns out that the "shear structure" of G_K can now be expressed in a particularly significant form, namely as a complex analytic structure for a closely related eight-dimensional manifold T. The interplay between complex analytic structure (i.e. "analyticity") and unitary structure (which, on the classical level becomes canonical, i.e. symplectic structure) seems to play an important role in modern theory of particles,[7, 12]

so there could be some significance in exhibiting such an interplay also at the level of a space-time analysis.

A second reason for desiring the specialization **to** $K = 0$ is that null geodesics are conformally invariant; that is to say, they depend only on the light cone structure of M and not on its metric. There are various reasons for believing that conformal invariance may actually have some basic role to play in physics, and that conformally invariant formalisms could have especial significance as "background formalisms" for physical theory. The notion of "causality" inasmuch as this refers the location of the light cones in space-time (and, therefore, to conformal structure) would seem to have a particular physical importance, more so than the actual space-time metric. In addition, all zero rest-mass free fields are, or can be made, conformally invariant.[3,5,10,11,28,30] This applies, in particular, to gravitation, if we interpret "free" to mean that we are considering linear theory, or at least the propagation of curvature in empty space (Bianchi identities)[30] rather than the non-linear response of curvature (or other fields) to curvature (Ricci identities). A number of interactions are also conformally invariant (e.g. electromagnetic interactions or the non-linear self-coupling in the "ϕ^4 scalar theory") but it is sometimes a little difficult to separate the conformal invariance of the "pure" interaction from the conformally non-invariant effects of the presence of rest-mass. It is even conceivable that all conformal invariance breaking is connected in some important way with the presence of mass. It is rest-mass which (apparently) is responsible for defining the scale of phenomena and hence the metric of space-time;[33,41] it provides us with the most obvious obstacle to a belief in a universal validity of conformal invariance in nature. The other most obvious place in which conformal invariance is broken is in gravitational interactions. Again it is mass which is involved in an essential (but now different) way. Indeed it is tempting to believe that there may be a common origin to these two aspects of conformal invariance breaking. In any case, to think of basic physical processes in terms of either conformal invariance, or the breaking of conformal invariance, seems to be a fruitful point of view. To this end, it is very useful to employ a formalism which makes this conformal invariance manifest wherever it *is* present.

The particular choice of formalism that I have in mind, namely *twistor theory*[31,34,36] is motivated partly by considerations of this kind. Twistors are, in fact, the spinors[4,6] for the fifteen parameter conformal group[16], which is the space-time symmetry group for the zero rest-mass free-field equations (including linearized gravitation). Any finite-dimensional representation of the conformal group is equivalent to a twistor representation. Also, infinite-dimensional representations can be conveniently described in terms of functions of twistors. Thus, twistors play a role with regard to the conformal group analogous to the role played by two-component spinors with regard to the Lorentz group[17] or rotation group. Now, the combinatorial model that I was discussing earlier for the description of non-relativistic angular momentum was based on the representations of the rotation group. In fact the particular combinatorial rules that I had in mind for evaluating the norm of a spin-network[35] were derived directly from the two-component spinor algebra. In a corresponding way we might expect to be able to derive a combinatorial calculus based on twistors, where the conformal group now takes over the role previously enjoyed by the rotation group.

It was, in fact, the hope of generalizing the spin-network model to make it more realistic which provided other parts of the motivation for the original introduction of twistor theory. The two most obvious respects in which the spin-network model is unrealistic are that it is a non-relativistic scheme and that there is no provision for the mixing of spin with orbital angular momentum. These two inadequacies are related to one another; indeed, they are both aspects of the fact that any relative velocity between the different units has been neglected. But it was always clear that the removal of these inadequacies in the model would involve considerably more than just the simple substitution of one group by another. Once the scheme is made relativistic then we must encounter some of the difficulties involved in passing from a quantum theory to a quantum field theory; once orbital angular momentum is brought in, then we have to contend not only with directions in space(-time) and angles, but also with locations and distances.

Accepting that there must be some essential new features arising, the twistors do seem to provide the right kind of generalization of the non-relativistic (i.e. SU(2)) two-component spinors which formed the basis for the spin-network theory. Twistors are genuinely spinorial objects and so can still handle half-odd spin values, they are completely relativistic (in the sense of special relativity) and they adequately mix together the concepts of angular and translational displacements (so that spin and orbital angular momenta will combine together in the appropriate way). The use of the conformal group - and hence the locally isomorphic "twistor group" SU(2,2) - rather than the Poincaré group, arises partly from technical mathematical reasons (connected with the semi-simplicity of SU(2,2)) and partly from reasons mentioned above concerning physical importance of conformal invariance. (In any case Poincaré invariance is easily extracted from a framework designed to handle conformal invariance.) The main essentially new feature which is involved arises from the fact that the conformal group (and SU(2,2)) possesses infinite-dimensional irreducible representations. Most particularly, the zero rest-mass free fields provide such representations. This implies that the twistor algebra must be employed more subtly than in the direct way in which the two-component spinor algebra generated the spin-network theory. Thus it may prove to be.

difficult to reduce the resulting twistor calculations (which at present involve contour integration) to a set of purely combinatorial rules. There seems to be nothing in principle against the possibility of doing this, however.

But how is twistor theory to be reconciled with general relativity? The conformal group refers only to symmetries of conformally flat space-time. Gravitation on the other hand implies the existence of conformal curvature. My point of view with regard to this question is really the one that I have been trying to stress throughout this article. Imagine that we have been able to develop the twistor theory to the point at which calculation can be expressed in terms of certain combinatorial rules. By analogy with the spin-network theory, we might expect that these rules would enable us to calculate the probability of occurrence of certain types of (graphically defined) situations in the universe. From these probabilities we should be able to extract geometrical concepts which would emerge as "well-defined" under suitable circumstances. This would then lead to a concept of local geometry which would be a Minkowskian character (assuming that conformal invariance breaking has been adequately incorporated into the theory - otherwise we should presumably only obtain a local conformal geometry). The Minkowski geometries that we extract "locally" might not be consistent with one another over the whole universe.

The concept of space-time "point" that we extract in one region would then appear to be "fuzzed out" from the point of view of the geometry defined in some other region.*
This "fuzzing out" would be of the nature of that obtainable by a unitary transformation in Hilbert space. On the classical level, this would appear as the result of a canonical transformation applied to a suitable space (namely twistor space - closely related to G_K with $K = 0$). From this space we do the best we can to extract a concept of space-time "point" which has some form of universal validity, but we find that having done this, the space-time that we finally construct is no longer conformally flat, the conformal curvature being directly relatable to this "fuzzing out" of points as "viewed" from distant regions.

To a considerable extent the above programme is, of course, speculation. Nevertheless, the present state of twistor theory does have a number of points of contact with it. To illustrate something of this, I should be more explicit about the nature of twistors. Let us, in the first instance, choose M to be Minkowski space-time. Choose an origin O and consider a classical special-relativistic system whose total momentum P_a is null and future-pointing and whose total angular momentum tensor M^{ab} has the property that the spin vector constructed from it (and P_a) is proportional to the momentum P_a :

---

* This is necessary because the local Minkowski geometries are not really tangent spaces. They have to merge one into the other when our point of view changes as we move around the universe. In a sense, they are all really the same "space".

$$P^a M^{bc} e_{abcd} = 2sP_d \ . \tag{10}$$

This is a normal requirement[13] for the momentum and angular momentum structure for a zero rest-mass particle (e_{abcd} being the alternating tensor). The quantity s is the spin-helicity; that is to say, $|s|$ is the intrinsic spin, while the sign of s measures the helicity. Translating into a two-spinor notation for P_a and M^{ab} we obtain.

$$P_{AA'} = \bar{\pi}_A \pi_{A'} \ , \tag{11}$$

for some $\pi_{A'}$, whence from (10)

$$M^{AA'BB'} = i\omega^{(A}\bar{\pi}^{B)}\varepsilon^{A'B'} - i\bar{\omega}^{(A'}\pi^{B')}\varepsilon^{AB} \tag{12}$$

for some ω^A (where the round brackets indicate symmetrization). The pair of spinors $(\omega^A, \pi_{A'})$ defines P_a and M^{ab} uniquely, while P_a and M^{ab} define $(\omega^A, \pi_{A'})$ up to the combined phase transformation

$$\omega^A \to e^{i\theta}\omega^A \ , \quad \pi_{A'} \to e^{i\theta}\pi_{A'} \qquad (13)$$

(θ real). Relative to the origin O , this pair of spinors represents a twistor[31,36] Z^α :

$$Z^\alpha \leftrightarrow (\omega^A, \pi_{A'}) \ . \qquad (14)$$

If we pass to a new origin $\tilde{O}$ whose position vector relative to O is h^a , then relative to $\tilde{O}$ we must represent the twistor Z^α by $(\tilde{\omega}^A, \tilde{\pi}_{A'})$ where

$$\tilde{\omega}^A = \omega^A - i\pi_{A'}h^{AA'} \ , \quad \tilde{\pi}_{A'} = \pi_{A'} \qquad (15)$$

this being consistent with the transformation of momentum and angular momentum:

$$\tilde{P}_a = P_a \ , \quad \tilde{M}^{ab} = M^{ab} - h^a p^b + p^a h^b \ . \qquad (16)$$

If we pass from the space-time metric g_{ab} to a conformally related metric $\hat{g}_{ab}$, according to the conformal rescaling[33]

$$\hat{g}_{ab} ,= \Omega^2 g_{ab} \ , \quad \hat{\varepsilon}_{AB} = \Omega\varepsilon_{AB} \ ; \qquad (17)$$

then Z^α must be represented by $(\hat{\omega}^A, \hat{\pi}_{A'})$ where

$$\hat{\omega}^{A} = \omega^{A} , \quad \hat{\pi}_{A'} = \hat{\pi}_{A'} + i\omega^{A}\Omega^{-1}\nabla_{AA'}\Omega . \qquad (18)$$

The two transformations (15) and (18) (and also any homogeneous Lorentz transformation) have the property that they are linear transformations (of unit determinant) which leave the form

$$Z^{\alpha}\bar{Z}_{\alpha} = \omega^{A}\bar{\pi}_{A} + \pi_{A'}\bar{\omega}^{A'} \qquad (19)$$

$$= +2s$$

invariant; where the complex conjugate $\bar{Z}_{\alpha}$ of the twistor Z^{α} is represented as

$$\bar{Z}_{\alpha} \leftrightarrow (\bar{\pi}_{A}, \bar{\omega}^{A'}) \qquad (20)$$

relative to the origin O. Since the Hermitian form (19) has signature (++--), such transformations generate the group SU(2,2) (this being 4 - 1 isomorphic with the connected component of the conformal group).

Twistors for which $Z^{\alpha}\bar{Z}_{\alpha} = 0$ are called null twistors. A null twistor Z^{α} defines the null straight line ζ in M which is the locus of points whose position vectors x^{a} satisfy

$$i\pi_{A'}x^{AA'} = \omega^{A} \tag{21}$$

(cf. (15)). We may think of ζ as *the world line of a zero rest-mass particle of momentum* p_a *and zero intrinsic spin*, this particle having angular momentum M^{ab} relative to 0 . The *non-null* twistors (i.e. $Z^{\alpha}\bar{Z}_{\alpha} \neq 0$) may be thought of as describing zero rest-mass particles with *non-zero intrinsic spin*, but they cannot be uniquely localized in terms of a single world-line in space-time.

Two null straight lines η and ζ will *intersect* (possibly at conformal infinity) if and only if their corresponding twistors Y^{α} and Z^{α} are *orthogonal*:

$$Y^{\alpha}\bar{Z}_{\alpha} = 0 ; \tag{22}$$

in which case the *light cone* whose vertex is the intersection q of η and ζ , is generated by null lines ξ described by null twistors

$$X^{\alpha} = \lambda Y^{\alpha} + \mu Z^{\alpha} . \tag{23}$$

Since any point q can be represented by its light cone, the linear set (23), denoted by $T(q)$, gives a twistor way of realizing the point q . If desired, we can use a twistor

$$Q^{\alpha\beta} = Y^{\alpha}Z^{\beta} - Z^{\alpha}Y^{\beta} \tag{24}$$

to represent T(q), and hence the point q. (Generally, twistors of higher valence can be constructed from twistors like Z^{α} or $\bar{Z}_{\alpha}$ by means of the usual tensor rules.)

According to a theorem due to Kerr[31], the condition that a congruence of null straight lines be shear-free can be stated very elegantly in twistor terms as the fact that the congruence be representable as the null solutions of an equation

$$\Phi(Z^{\alpha}) = 0 \tag{25}$$

(or by the limiting case of such a construction) where Φ is a complex analytic (holomorphic) function of the components of the twistor Z^{α}. Thus, the shear-free condition is interpreted, in twistor terms, as essentially the Cauchy-Riemann equations $\partial\Phi/\partial\bar{Z}_{\alpha} = 0$. In other words, the "shear-structure" for the space T of twistors Z^{α}, may be thought of as a complex (analytic) structure on T. We can use this structure to locate the T(q) manifolds, and hence the points of M (since light cones are the only non-shearing null hyperfurfaces in M, apart from null hyperplanes).

Let us turn to the case when M is a (globally hyperbolic) curved space-time. The concept of a null twistor Z can be adapted without difficulty from the Minkowski case: we simply interpret Z as a null geodesic ζ in M at each of whose points is a spinor $\pi_{A'}$, parallelly propagated along ζ, such that the "momentum vector" P_a, given by $P_{AA'} = \bar{\pi}_A \pi_{A'}$ is tangent to ζ. On the other hand, there appears to be no way of uniquely associating a non-null twistor with some well-defined structure on M. However it is useful to postulate the existence of a space of non-null twistors, into which the space T_0 of null twistors is to be embedded as a hypersurface. This gives us an eight-dimensional manifold T, the space T_0 being a seven-dimensional submanifold.

The space T is supposed to possess a symplectic structure, namely a real two-form, which I shall still denote by ω, this being the exterior derivative $\omega = d\phi$ of a complex one-form ϕ on T. The imaginary part of ϕ is to be the exterior derivative of a scaler field $+s$ on T, the hypersurface T_0 being defined by $s = 0$. If a region of T_0 refers to null geodesics which enter some region of flat space in M, then we can use the representation of twistors given in (14), etc. and be more explicit as to the definition of these forms. We can set[34]

$$\omega = i\,dZ^\alpha \wedge d\bar{Z}_\alpha, \quad \phi = iZ^\alpha d\bar{Z}_\alpha \text{ and } s = +\tfrac{1}{2}Z^\alpha \bar{Z}_\alpha \qquad (26)$$

(cf. (19)). In the case when $s = 0$ (vanishing of intrinsic spin) we can substitute $Z^{\alpha} \leftrightarrow (ix^{AA}\pi_{A'}, \pi_{A'})$ (cf. (21)) into the right-hand sides and verify directly the formal equivalence with (3) and (4). If the null geodesics pass theough *two* regions of flat space, then the definitions (26) arising from each will agree with one another.[34] We can also use (26) if the null geodesics do *not* enter any region of flat space, provided we are concerned only with T_0 and interpret the dZ^{α}'s suitably.

The space G_0 (i.e. G_K with $K = 0$) of null geodesics, that we considered earlier, can be regarded as obtainable from T by a process analogous to that by which G_0 was previously obtained from the tangent bundle C. The (squared) "Hamiltonian" is now the quantity $s = +\frac{1}{2}Z^{\alpha}\bar{Z}_{\alpha}$. We are concerned with the seven-dimensional submanifold T_0 of T given by $s = 0$. The vector field corresponding to χ is now

$$-\frac{i}{2}Z^{\alpha}\frac{\partial}{\partial Z^{\alpha}} + \frac{i}{2}\bar{Z}_{\alpha}\frac{\partial}{\partial \bar{Z}_{\alpha}} \tag{27}$$

which generates the transformations $Z^{\alpha} \to e^{i\theta}Z^{\alpha}$ (cf. (13)). We factor out by these phase transformations to pass from T_0 to the six-dimensional space G_0. (See figure 9.) This process enables us to carry over the symplectic structure ω, and also ϕ, from T, now, to the space G_0. The result agrees with the previous construction[9]. Thus, the symplectic structure of T can, when restricted to T_0, be given some real geometrical significance.

CHAPTER 56

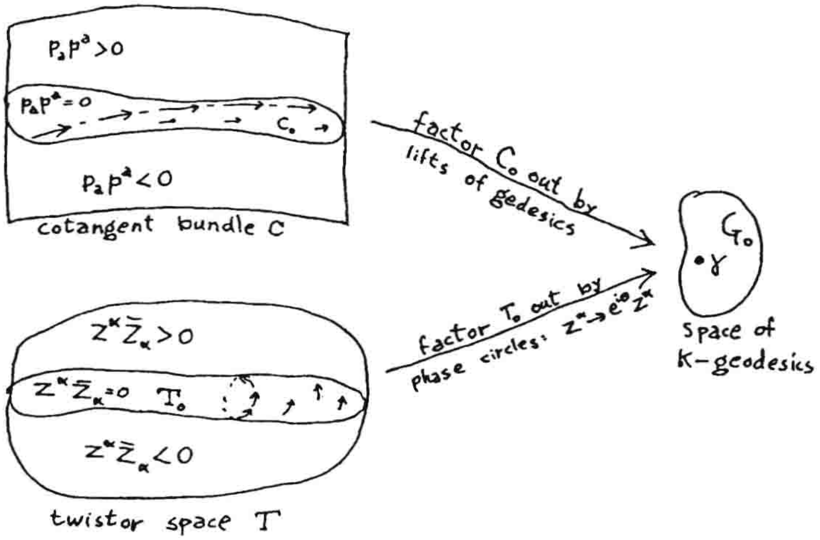

$P_a P^a > 0$
$P_a P^a = 0$
C_0
$P_a P^a < 0$
cotangent bundle C
factor C_0 out by lifts of gedesics
G_0
γ
Space of K-geodesics
factor T_0 out by phase circles: $Z^\alpha \to e^{i\theta} Z^\alpha$
$Z^\alpha \bar{Z}_\alpha > 0$
$Z^\alpha \bar{Z}_\alpha = 0$
T_0
$Z^\alpha \bar{Z}_\alpha < 0$
twistor space T

FIGURE 9.

The symplectic structure of G_0 arises from that of C or from that of T in a similar way.

It is actually possible to go considerably further than this in explicitly exhibiting the form of canonical transformation on T , which is induced, for example, by the presence of a gravitational wave in M . An interplay between analytic functions and the symplectic structure of T appears again in a surprising way[36] I do not want to go into all this here. The main point I wanted to make is that conformal curvature shows up classically on the space T in terms of canonical transformations $Z^\alpha \to \tilde{Z}^\alpha(Z^\beta, \bar{Z}_\gamma)$, the form ω being preserved[34]. Such a transformation shifts the complex structure of T . It is the complex structure of T (being its "shear structure") which serves to "locate" the points in M . Such a shift causes "good" null cones to be transformed into cones which do not focus cleanly at a proper vertex. In this way it has the effect of "fuzzing out" points which had previously appeared to be "good" points and vice versa. Quantum mechanically, we would expect this effect to result from a unitary transformation $Z^\alpha \to Z^\alpha(Z^\beta, \partial/\partial Z^\delta)$.

In my descriptions up to now I have essentially ignored the question of the topology of M. I can, however, see no real objection to applying these ideas to quite general space-time manifolds with complicated topology. The restriction to a globally hyperbolic M allows one to avoid the more serious problems which might arise. But it also precludes any possiblity if having a space-time with a changing topology for example. (Global hyperbolicity is equivalent[18] to the existence of a Cauchy hypersurface for M.) If global hyperbolicity is dropped, then T_0 and G_0 could exhibit certain pathologies. If M is assumed to be strongly causal,[19,20,33] then the worst of these pathologies can be avoided. But there still remains the possibility that T_0 and G_0 could turn out to be non-Hausdorff manifolds under such circumstances.

Finally, I should say something about the status of the details of twistor theory at the moment[36]. The theory allows computations to be carried out using contour integrals, which are intended to represent (cross sections for) scattering processes involving electromagnetic and gravitational interactions etc. (It is such computations as these which one would hope to be able ultimately to re-express in a purely combinatorial form.) At the time of writing, there is some evidence that those computations which represent the conformally invariant part of the theory (e.g. electromagnetic interactions without rest-mass) are substantially correct. On the other hand it has not yet been possible to find the correct way of handling the conformally non-invariant parts of the theory (e.g. rest-mass and gravitational interactions). But there appears to be no insuperable obstacle eventually to achieving this.

On the nature of quantum geometry. *Magic without magic* (J. A. Wheeler Festschrift), pp. 333–54, Freeman, 1973

REFERENCES

1. Aharonov, Y. (ref. to be completed)

2. Ahmavaara, Y., J. Math. Phys. 6, 87(1965).

3. Bateman, H., Proc. Lond. Math. Soc. 8, 223(1910)

4. Brauer, R. & Weyl, H., Am. J. Math. 57, 425(1935)

5. Buchdahl, H. A., Nuovo Cimento 11, 496(1959)

6. Cartan, E., Ann. Sci. Ecole Norm. Sup. 31, 263(1914)

7. Chew, G. F., S-Matrix Theory of Strong Interactions (W. A. Benjamin, Inc., New York, 196).

8. Choquet-Bruhat, Y., in Battelle Rencontres C. M. DeWitt & J. A. Wheeler, Eds. (W. A. Benjamin Inc., New York, 1968), Chap. IV.

9. Crampin, M. & Pirani, F. A. E., (Rosen Festschrift, 1970).

10. Cunningham, E., Proc. Lond. Math. Soc. 8, 77(1910)

11. Dirac, P. A. M., Ann. Math. 37 429(1936).

12. Eden, R. J., Landshoff, P. V., Olive, D.I. & Polkinghorn, J.C. The Analytic S-Matrix (Cambridge University Press).

13. Fierz, M., Helv. Phys. Acta 13, 95(1940)

14. Finkelstein, D. (ref. to be completed)

15. Flanders, H., Differential Forms, with Applications to the Physical Sciences (Academic Press, New York, 1963)

16. Fulton, T., Rohrlich, F. & Witten, L., Rev. Mod. Phys. 34, 442(1962).

17. Gel'fand, I. M., Graev, M. I. & Vilenkin, N. Ya, Generalized functions V: Integral Geometry and representation Theory (Academic Press, New York, 1966).

18. Geroch, R., J. Math. Phys. 11, 437(1970)

19. Hawking, S. W. Proc. Roy. Soc. (Lond.) A300 187(1967)

20. Hawking, S. W. & Penrose, R., Proc. Roy. Soc. (Lond.) A314, 529(1970).

21. Helgason, S., Differential Geometry & Symmetric Spaces (Academic Press, New York, 1962).

22. Hermann, R. (ref to be completed).

23. Hicks, N. J. Notes on Differential Geometry (D. van Nostrand Co. Inc., Princeton, 1965).

24. Kronheimer, E. H. & Penrose, R., Proc. Camb. Phil. Soc. 63, 481(1967).

25. Laugwitz, D., Sitz. Bayer. Akad. Wiss. 4, 41(1959)

26. Lichnerowicz, A. in Battelle Rencontres C. M. DeWitt & J, A. Wheeler, Eds. (W. A. Benjamin Inc., New York, 1968) Chap. V.

27. Mackey, G. W. Mathematical Foundations of Quantum Mechanics (W. A. Banjamin Inc. New York, 1963).

28. McLennan, J. A.Jr., Nuovo Cimento 10, 1360(1956).

29. Menger, K. (ref to be completed).

30. Penrose, R., Proc. Roy. Soc. (Lond.) A284, 159(1965).

31. Penrose, R., J. Math. Phys. 8, 345(1967)

32. Penrose, R., An Analysis of the structure of space-time (Adams Prize Essay, Cambridge University 1966).

33. Pensose, R., in Battelle Rencontres C. M. DeWitt & J. A. Wheeler, Eds. (W. A. Benjamin Inc. New York, 1968) Chap VII

34. Penrose, R., Int. J. Theor. Phys. 1, 61(1968)

35. Penrose, R., in Quantum Theory and Beyond, E. T. Bastin, Ed. (Cambridge University Press, 1970).

36. Penrose, R., (ref to be completed).

37. Robinson, A., Proc. Nederl. Akad. Wetensch 64, (1961).

38. Schild, A., Phys. Rev. 73, 414(1948).

39. Schmieden, C. & Langwitz, D., Math Zeitschr. 69, 1(1958)

40. Schrodinger, E., Science and Humanism (Cambridge University Press, 1956).

41. Synge, J. L., Relativity, the Special Theory (North Holland).

57

INTERNAL INSTABILITY IN A REISSNER–NORDSTRÖM BLACK HOLE

Int. J. Th. Phys. **7** (1973), 183–97. With M. Simpson

International Journal of Theoretical Physics, Vol. 7, No. 3 (1973), pp. 183–197

Internal Instability in a Reissner–Nordström Black Hole

MICHAEL SIMPSON and ROGER PENROSE

Department of Mathematics, Birkbeck College, London

Received: 1 *April* 1972

Abstract

The question of the effect of asymmetries in gravitational collapse is investigated by considerations of test electromagnetic fields in an extended Reissner–Nordström background. It is found, with the aid of computer calculations, that instabilities in the test field arise at the inner (Cauchy or anti-event) horizon, though not at the outer (event) horizon. Thus it is reasonable to infer that in the full coupled Einstein–Maxwell theory the inner horizon will not survive as a non-singular hypersurface when asymmetric perturbations are present, but will instead become a space-time curvature singularity.

The gravitational collapse of a star which is too massive to form a white dwarf or neutron star presents a now familiar picture. If spherical symmetry is assumed, then collapse through a black hole to a central space-time singularity at which curvatures mount to infinity is implied by general relativity (Penrose, 1969a). It is also known from general theorems (Hawking & Penrose, 1970) that even when spherical symmetry is not assumed, a space-time singularity must nevertheless still arise whenever trapped surfaces occur. However, the nature and location of these singularities is left completely open by the theorems. According to a frequently stated conjecture (Carter, 1971; Israel, 1967, 1968; Hawking, 1972), the external field of the black hole which results from a gravitational collapse should approach that of a Kerr–Newman (Kerr, 1963; Boyer & Lindquist, 1967; Newman *et al.*, 1965) solution of the Einstein–Maxwell equations, characterized by just three parameters: mass m, charge e, and angular momentum a. There is however no reason to believe that the internal field near the (ring) singularity of these solutions should be an accurate representation of the physical space-time resulting from a realistic collapse. Indeed, the presence of closed timelike curves near the singularities (Carter, 1968) would seem to argue against any too close relation between the models and physical reality.

The Kerr–Newman solutions with $m > \sqrt{(a^2 + e^2)} > 0$ possess the

Copyright © 1973 Plenum Publishing Company Limited. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission of Plenum Publishing Company Limited.

characteristic feature that there is not only an outer (absolute) event horizon (surface of the black hole) but also an inner horizon (the anti-event horizon) (Penrose, 1969b) which is a Cauchy horizon (Hawking, 1966, 1967) for any appropriate initial data hypersurface used to set up the problem. Our contention in this note is that if the initial data is generically perturbed then the Cauchy horizon does not survive as a non-singular hypersurface. It is strongly implied that instead, genuine space-time singularities will appear along the region which would otherwise have been the Cauchy horizon.

We consider here only a special case, namely that when the angular momentum is zero, but mass and charge parameters m and e are both present (with $m > |e|$). We add an electromagnetic test field as a perturbation and show (by computer calculation) that singularities in the electromagnetic field occur at the Cauchy horizon. It is reasonable to infer that when the gravitational–electromagnetic coupling is added, then the Cauchy horizon would degenerate into a curvature singularity.

Our unperturbed space-time can be described by the Reissner–Nordström metric, given in the form†

$$\begin{aligned} ds^2 = (1 - 2m/r + e^2/r^2)\,dt^2 - (1 - 2m/r + e^2/r^2)^{-1}\,dr^2 \\ - r^2(d\theta^2 + \sin^2\theta\,d\phi^2) \end{aligned} \tag{1}$$

where r, t, θ, ϕ are the radial, time, and two angular coordinates respectively. The coordinate singularities due to the distinct zeros r_+, $r_-(r_+ > r_-)$ of $(1 - 2m/r + e^2/r^2)$, corresponding to the event horizon and Cauchy horizon respectively, divide the manifold covered by this coordinate system into the three regions $r > r_+$, $r_- < r < r_+$ and $r < r_-$. Each of these can be parametrised by the retarded and advanced null coordinates u, v given by

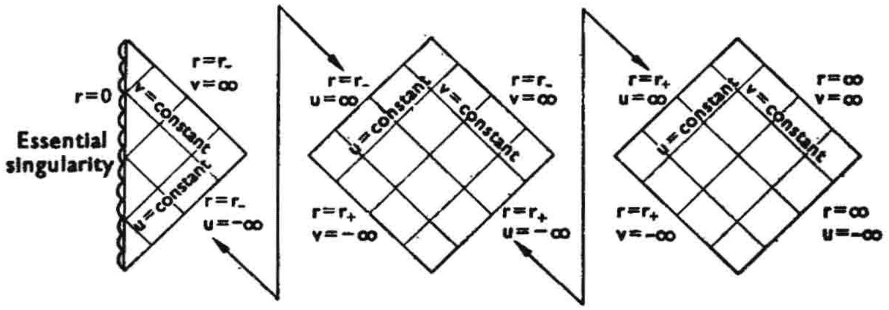
$$\begin{aligned} u &= \begin{cases} t - f(r), & r > r_+ \text{ and } r < r_- \\ t + f(r), & r_- < r < r_+ \end{cases} \\ v &= \begin{cases} t + f(r), & r > r_+ \text{ and } r < r_- \\ -t + f(r), & r_- < r < r_+ \end{cases} \end{aligned} \tag{2}$$

where $f(r) = \int (1 - 2m/r + e^2/r^2)^{-1}\,dr$ and u, v have range $(-\infty, \infty)$ in each of the three regions. The parts of the manifold can be represented by the 'blocks' of Fig. 1 for constant values of the angular coordinates. The edges of the blocks are identified as indicated. To the three regions A, B, C there correspond further regions A', B', C' obtained by applying the transformation $u \to -u$, $v \to -v$ which reverses the light cone structure. A maximal analytic extension‡ of the Reissner–Nordström solution may then

---

† See, for example, C. Møller (1952). *The Theory of Relativity*. Clarendon Press, Oxford.

‡ This procedure is due to J. C. Graves and D. R. Brill (1960). *Physical Review*, **120**,


r=0
Essential singularity
r=r₋
v=∞
v=constant
u=constant
r=r₋
u=−∞
r=r₋
u=∞
r=r₋
v=∞
r=r₊
v=−∞
r=r₊
u=−∞
r=r₊
u=∞
r=∞
v=∞
r=r₊
v=−∞
r=∞
u=−∞

(a) Region A: $r<r_-$

(b) Region B: $r_-<r<r_+$

(c) Region C: $r>r_+$

Figure 1.

be constructed by piecing together copies of the six blocks in an analytic fashion so that each overlapping edge is covered by either a (u,r) or (v,r) system of coordinates (with the exception of the corners of each block). The resulting 'ladder' shown in Fig. 2 is infinitely extendible in both directions.

In this space-time the edge of a freely collapsing body is represented by a future-directed timelike geodesic which passes from the exterior region *C* through *B* to the interior region *A*. The nature of the analytic extension allows us to contemplate the possibility (Penrose, 1968) of an observer following in the wake of the collapsing body from *C* to *B*, then avoiding

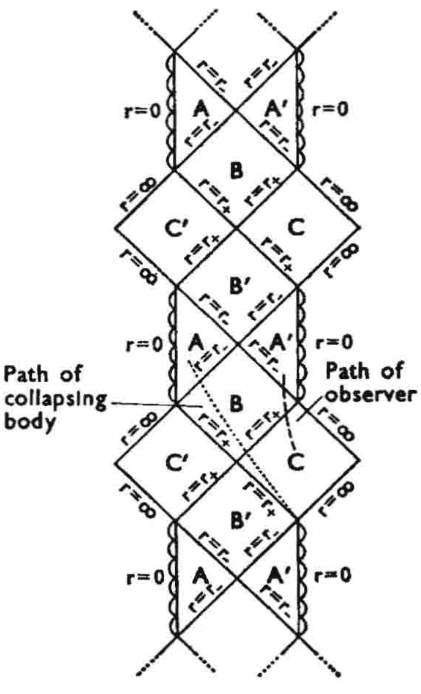

r=0
A
A′
r=0
B
C′
C
B′
r=0
A
A′
r=0
Path of collapsing body
Path of observer
B
C′
C
B′
r=0
A
A′
r=0
r=r₋
r=r₊
r=∞

Figure 2.

region A by pursuing instead a (necessarily) non-geodesic trajectory through $r=r_-$ into region A', whose history is not determined by data specified in C. The observer is not however assured of a smooth passage over the Cauchy horizon $r=r_-$ between B and A'. In the instant before crossing he is in a position to observe the entire history of region C and it may be that in consequence of receiving the full (back-scattered) effects of the outgoing radiation field from C he will encounter unbounded curvatures in the vicinity of $r=r_-$. If arbitrary smooth initial data for our electromagnetic test field given on a spacelike hypersurface in C fail to determine limiting values as $r=r_-$ is approached, things look bad for our observer. If such limits are forthcoming, he appears to have chances of getting

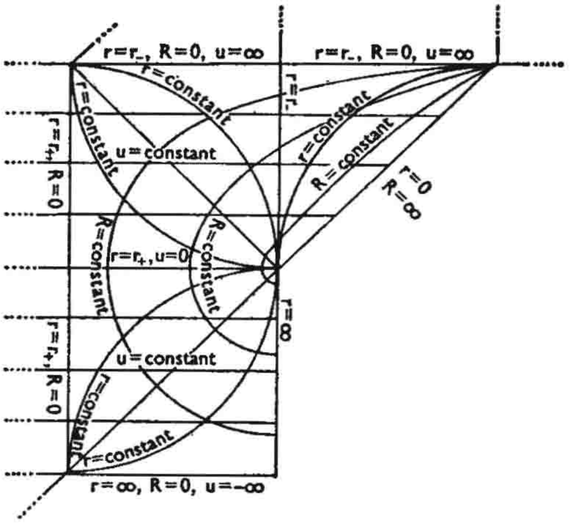


Figure 3.

through. We proceed to set up the necessary apparatus for making such an investigation.

After first performing the conformal transformation $ds \to r^{-1} ds$ we further transform the metric, leaving the angular coordinates unchanged and defining new coordinates u' and R by

$$u' = \begin{cases} -e^{-Au}, & r_+ \leqslant r < \infty \\ e^{Au}, & 0 \leqslant r \leqslant r_+ \end{cases}$$

$$R = \frac{r_+ - r}{Au' r r_+}, \qquad 0 \leqslant r < \infty \qquad \text{where } A = \frac{r_+ - r_-}{2r_+^{2}} \tag{3}$$

These coordinates cover the three regions C, B, and A'. The outgoing null directions are given by $u' = \text{constant}$, while R is an affine parameter along those directions, i.e. $\nabla^a u' \nabla_a R = 1$. Dropping the prime on u, Fig. 3 shows the disposition of the coordinate curves. The metric is now given by

$$ds^2 = g(u, R)\, du^2 + 2du\, dR - (d\theta^2 + \sin^2\theta\, d\phi^2) \tag{4}$$

where

$$g(u,R) = R^2\{(3\alpha-2) + \tfrac{1}{2}(1-\alpha)(3\alpha-1)uR + \tfrac{1}{4}\alpha(1-\alpha)^2 u^2 R^2\}$$

and

$$\alpha = \frac{r_-}{r_+}, \qquad 0 \leqslant \alpha < 1$$

$\alpha = 0$ gives the Schwarzschild limiting case.

Seeking to employ the Newman–Penrose spin coefficient formalism (Newman & Penrose, 1962) we put $l^a = \nabla^a u$, tangent to the outgoing null hypersurfaces, and define three other vectors n^a, m^a, $\bar{m}^a$ to complete a null tetrad satisfying the conditions that n^a shall be a real future-pointing null vector lying in the plane spanned by $\nabla^a u$ and $\nabla^a R$, m^a and its conjugate $\bar{m}^a$ shall be complex null vectors, and $l_a n^a = -m_a \bar{m}^a = 1$, $l_a m^a = n_a m^a = 0$. We take

$$\begin{aligned} n^a &= \nabla^a R + \tfrac{1}{2} g(u,R)\,\nabla^a u \\ m^a &= -\frac{1}{\sqrt{2}}(\nabla^a \theta + i \sin\theta \nabla^a \phi) \end{aligned} \tag{5}$$

Using this tetrad, nine of the twelve spin coefficients are zero, the exceptions being

$$\begin{aligned} \alpha &= \tfrac{1}{2}(n^a \bar{m}^b \nabla_b l_a - \bar{m}^a \bar{m}^b \nabla_b m_a) = -\frac{\cot\theta}{2\sqrt{2}} \\ \beta &= \tfrac{1}{2}(n^a m^b \nabla_b l_a - \bar{m}^a m^b \nabla_b m_a) = \frac{\cot\theta}{2\sqrt{2}} \\ \gamma &= \tfrac{1}{2}(n^a n^b \nabla_b l_a - \bar{m}^a n^b \nabla_b m_a) = \frac{1}{4}\frac{\partial g}{\partial r} \end{aligned} \tag{6}$$

where $g(u,R)$ is given by (4).

Defining the three complex components of the electromagnetic field tensor F_{ab} by

$$\begin{aligned} \phi_0 &= F_{ab}\, l^a m^b \\ \phi_1 &= \tfrac{1}{2}F_{ab}(l^a n^b - \bar{m}^a m^b) \\ \phi_2 &= F_{ab}\, \bar{m}^a n^b \end{aligned} \tag{7}$$

and writing $D = l^a \nabla_a$, $\Delta = n^a \nabla_a$, $\delta = m^a \nabla_a$, the spin coefficient form of Maxwell's equations is

$$\begin{aligned} D\phi_1 - \bar{\delta}\phi_0 &= -2\alpha\phi_0 \\ D\phi_2 - \bar{\delta}\phi_1 &= 0 \\ \delta\phi_1 - \Delta\phi_0 &= -2\gamma\phi_0 \\ \delta\phi_2 - \Delta\phi_1 &= -2\beta\phi_2 \end{aligned} \tag{8}$$

We next introduce (Penrose, 1967; Newman & Penrose, 1968) the angular differential operator $\eth$. A quantity η is said to have spin weight s if under a transformation $m^\mu \to e^{i\chi} m^\mu$ it transforms as $\eta \to e^{is\chi}\eta$. It is readily seen from (7) that ϕ_0, ϕ_1, ϕ_2 have spin weights 1, 0 and -1 respectively; $\eth$ is then defined for a quantity η of spin weight s by the equation

$$\eth\eta = -(\sin\theta)^s \left\{\frac{\partial}{\partial\theta} + \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\}\{(\sin\theta)^{-s}\eta\} \tag{9}$$

The operator $\bar{\eth}$ is similarly defined:

$$\bar{\eth}\eta = -(\sin\theta)^{-s} \left\{\frac{\partial}{\partial\theta} - \frac{i}{\sin\theta}\frac{\partial}{\partial\phi}\right\}\{(\sin\theta)^{s}\eta\} \tag{10}$$

(9) and (10) reduce to

$$\begin{aligned} \eth\eta &= -\surd 2\,\delta\eta + s\eta\cot\theta \\ \bar{\eth}\eta &= -\surd 2\,\bar{\delta}\eta - s\eta\cot\theta \end{aligned} \tag{11}$$

Substituting ϕ_0, ϕ_1, ϕ_2 in turn for η in (11) and putting s equal to the corresponding spin weight, we see that Maxwell's equations (8) may be put in the form

$$\begin{aligned} D\phi_1 + \frac{1}{\surd 2}\bar{\eth}\phi_0 &= 0 \\ D\phi_2 + \frac{1}{\surd 2}\bar{\eth}\phi_1 &= 0 \\ (\Delta - 2\gamma)\,\phi_0 + \frac{1}{\surd 2}\eth\phi_1 &= 0 \\ \Delta\phi_1 + \frac{1}{\surd 2}\eth\phi_2 &= 0 \end{aligned} \tag{12}$$

We define the quantities

$$H_m = \int {}_1\bar{Y}_{1,m}\,\phi_0\,d\theta\sin\theta\,d\phi, \qquad m = -1, 0, 1 \tag{13}$$

where ${}_1\bar{Y}_{1,m}$ is a spin s spherical harmonic given by substituting for s and l in the defining equations

$$\begin{aligned} {}_sY_{l,m} &= \begin{cases} \left[\dfrac{(l-s)!}{(l+s)!}\right]^{1/2} \eth^s Y_{l,m}, & 0 \leqslant s \leqslant l \\ (-1)^s \left[\dfrac{(l+s)!}{(l-s)!}\right]^{1/2} \bar{\eth}^{-s} Y_{l,m}, & -l \leqslant s \leqslant 0 \end{cases} \\ {}_s\bar{Y}_{l,m} &= (-1)^{m-s}\,{}_{-s}Y_{l,-m} \end{aligned} \tag{14}$$

where $l = 0, 1, \ldots$; $m = -l, \ldots, l$; and $Y_{l,m}$ are the ordinary spherical harmonics. We note the property that if ρ is a quantity of spin weight $l+1$ then

$$\int {}_s\bar{Y}_{l,m}\, \bar{\eth}^{l-s+1}\, \rho\, d\theta \sin\theta\, d\phi = 0 \tag{15}$$

Dropping the subscript from H and using (11), (12) and (15), together with the three commutator relations for scalars ϕ

$$\begin{aligned} (\Delta D - D\Delta)\,\phi &= 2\gamma D\phi \\ (\delta D - D\delta)\,\phi &= 0 \\ (\bar{\eth}\eth - \eth\bar{\eth})\,\phi &= 2s\phi \end{aligned} \tag{16}$$

where in the last equation ϕ is considered to be of spin weight s, we derive a linear homogeneous second-order partial differential equation in H of hyperbolic type:

$$\mathscr{L}[H] = \{(\Delta - 4\gamma)\, D + (1 - 2D\gamma)\}\, H = 0 \tag{17}$$

Substituting for Δ, D, γ, we have

$$\mathscr{L}[H] = \left\{\frac{\partial^2}{\partial u\,\partial R} - \frac{g}{2}\frac{\partial^2}{\partial R^2} - \frac{\partial g}{\partial R}\frac{\partial}{\partial R} + \left(1 - \frac{1}{2}\frac{\partial^2 g}{\partial R^2}\right)\right\} H = 0 \tag{18}$$

or

$$\frac{\partial^2 H}{\partial u\,\partial R} - \frac{1}{2}\frac{\partial^2}{\partial R^2}(gH) + H = 0$$

where $g(u, R)$ is given by (4).

The canonical form of (18),

$$\left\{\frac{\partial^2}{\partial u\,\partial v} + A(u,v)\frac{\partial}{\partial v} + B(u,v)\right\} H = 0$$

requires for its explicit description the inverse (for R) of the transformation

$$v = \frac{R \exp\left[\dfrac{2(1-\alpha)}{2 + (1-\alpha)\,uR}\right]}{(2 - \alpha uR)^{\alpha^2}\,(2 + (1-\alpha)\,uR)^{(1-\alpha^2)}} \tag{19}$$

which is not forthcoming. Nor does Riemann's method of solution of the initial value problem for partial differential equations of hyperbolic type assist us since the adjoint equation to (18)

$$\mathscr{L}^*[J] = \frac{\partial^2 J}{\partial u\,\partial R} - \tfrac{1}{2} g \frac{\partial^2 J}{\partial R^2} + J = 0 \tag{20}$$

is related to (18) by

$$\frac{\partial^2}{\partial R^2}\mathscr{L}^*[X] = \mathscr{L}\left[\frac{\partial^2 X}{\partial R^2}\right] \tag{21}$$

i.e. if $J = X$ is a solution of $\mathscr{L}^*[J] = 0$ then $H = \partial^2 X/\partial R^2$ is a solution of $\mathscr{L}[H] = 0$. Finding exact solutions of (20), which must be achieved in order to construct a Green's function, is consequently equivalent to finding them for (18), and excepting one trivial case no exact solutions of either have been found.

We note that at $r = \infty$, given by $uR = 2/(\alpha - 1)$ in our present coordinate system, $1 - \frac{1}{2}\partial^2 g/\partial R^2 = 0$ and hence (18) reduces to

$$\left\{\frac{d}{du} + R(1 - \alpha)\right\} DH = 0$$

giving

$$\frac{d}{du}(R^2\, DH) = 0 \tag{22}$$

Thus from (13) we have the integral conservation law

$$\frac{d}{du}\int {}_1\bar{Y}_{1,m} R^2\, D\phi_0\, d\theta \sin\theta\, d\phi = 0 \tag{23}$$

the integration being carried out over the sphere at infinity. This integral is one of the Newman–Penrose absolutely conserved quantities (Penrose, 1967; Newman & Penrose, 1968).

In the apparent absence of explicit analytic solutions of (18) numerical solutions have been sought by the usual method for hyperbolic partial differential equations (Smith, 1969). The characteristic directions (null lines) are given by

$$\left.\begin{aligned} du &= 0 \\ \frac{dR}{du} + \tfrac{1}{2}g(u, R) &= 0 \end{aligned}\right. \tag{24}$$

and putting

$$p = \frac{\partial H}{\partial u}, \qquad q = \frac{\partial H}{\partial R} \qquad \text{and} \qquad f(u, R, q, H) = q\frac{\partial g}{\partial R} + H\left(\frac{1}{2}\frac{\partial^2 g}{\partial R^2} - 1\right)$$

the differential relationships

$$\begin{aligned} dp &= \tfrac{1}{2}g\, dq + f\, dR \\ dq &= f\, du \end{aligned} \tag{25}$$

represent (18) along the respective characteristics. We also have along $du = 0$,

$$dH = q\, dR \tag{26}$$

Successive approximations to H, p, q at any point X of the characteristic grid in terms of their values at immediately preceding grid points on the characteristics through X are obtained by iterative use of the finite difference equivalents of (25) and (26). If values of p are not required the first equation in (25) becomes redundant since p appears neither in the second of (25) nor in (26).

The solution of the second equation in (24) is given by (19). The R-coordinate of a point on the characteristic grid given by specified values of u, v could therefore be found using Newton's method to solve (19) to any desired accuracy, but this is a lengthy procedure to undertake for every point of the grid and is consequently employed only at points on the initial hypersurface. Elsewhere a Runge–Kutta approximation† using (24) gives the values of R at successive grid points on each v-characteristic. When $u = 0$ (19) becomes $v/R = \text{constant}$, and by scaling v so that this constant is unity we obtain a direct check on the accuracy of the approximation for R in the region $u \leqslant 0$.

Initial values of H, q are specified on a pair of null lines $u = u_0$, $v = v_0$ ($u_0 < 0$, $v_0 \geqslant 0$), the initial value of p if required being then uniquely determined on these characteristics. For the computation‡ it is necessary to choose a numerical value for the ratio $\alpha = r_-/r_+$ and we put $\alpha = \frac{2}{3}$, for which the maximum divergence of the v-characteristics occurs on $u = 0$. Along $v = \text{constant}$, dR/du has a maximum and minimum of order v^2 on either side of $u = 0$. The number of grid points required to furnish a reasonably accurate solution therefore rises sharply as v increases placing practical limits on the extent of the solution domain in this direction. Within these bounds results of interest have been obtained.

Several functions have been used as initial data. In particular setting $v_0 = 0$ and choosing H, p, q to be zero on $v = v_0$, the C^1 initial function

$$H_0 = \begin{cases} R^2(a-R)^2, & 0 \leqslant R \leqslant a \\ 0, & a \leqslant R \leqslant \dfrac{2}{(\alpha - 1)\,u_0} \end{cases} \tag{27}$$

for $u = u_0$ has been tried with various values of u_0 and a. Profiles along $u = \text{constant}$ of the solution obtained from (27) with $u_0 = -100$ and $a = 0{\cdot}0525$ are shown in Fig. 5 while cross-sections along $v = \text{constant}$ are displayed in Fig. 6. Figure 4 shows the relative disposition of the u- and v-characteristics employed, drawn to the same scale. These graphs show features typical of all the numerical solutions investigated.

By choosing the initial value of DH on the boundary to be zero we are assured by (22) of its remaining zero all the way up $r = \infty$. Graphs of H for any constant negative value of u are asymptotically flat in consequence. As u increases from -100 to 0 the number of oscillations of the solution for $u = \text{constant}$ goes up, remaining finite for any particular $u < 0$. On $u = 0$ we apparently have an unbounded number of well-damped oscillations, and as we enter the region $u > 0$ the amplitude of these oscillations rises sharply and they appear to diverge more and more violently as u

---

† See, for example, P. K. Henrici (1968). *Discrete Variable Methods in Ordinary Differential Equations*. Wiley, New York.

‡ All programs were run on the University of London's Atlas computer.

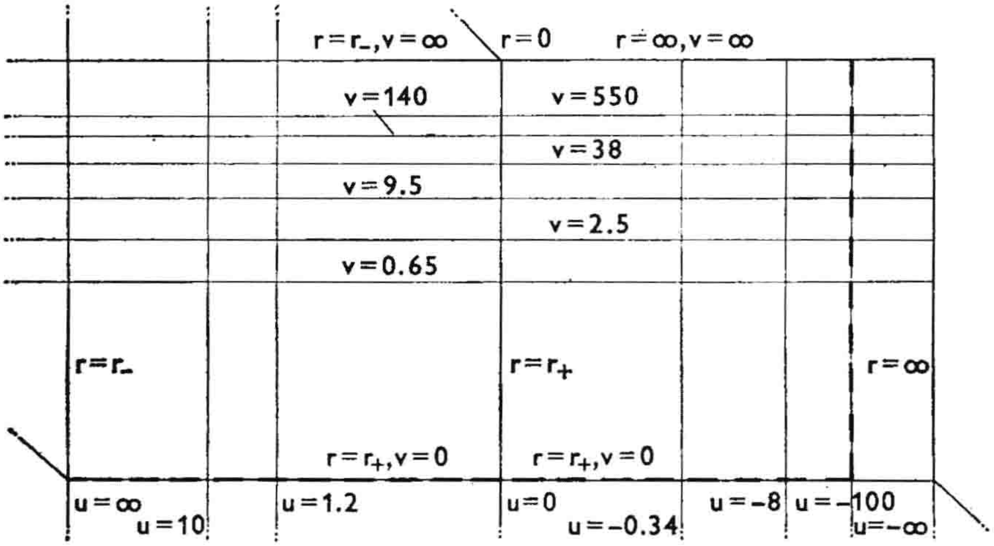

r=r₋,v=∞
r=0
r=∞,v=∞
v=140
v=550
v=38
v=9.5
v=2.5
v=0.65
r=r₋
r=r₊
r=∞
r=r₊,v=0
r=r₊,v=0
u=∞
u=10
u=1.2
u=0
u=−0.34
u=−8
u=−100
u=−∞

Figure 4.

further increases. It can be seen in graphs (e) and (f) of Fig. 5 that these unstable oscillations are situated 'very close' to the Cauchy horizon $r = r_-$ in terms of the affine parameter R.

The solution for $v =$ constant to the left of any neighbourhood of $u = 0$ appears to settle down rapidly as v is increased while the number of oscillations of decreasing amplitude near $u = 0$ continues to rise. It seems reasonable to extrapolate from this to the picture at $v = \infty$ for $u < 0$ featuring an

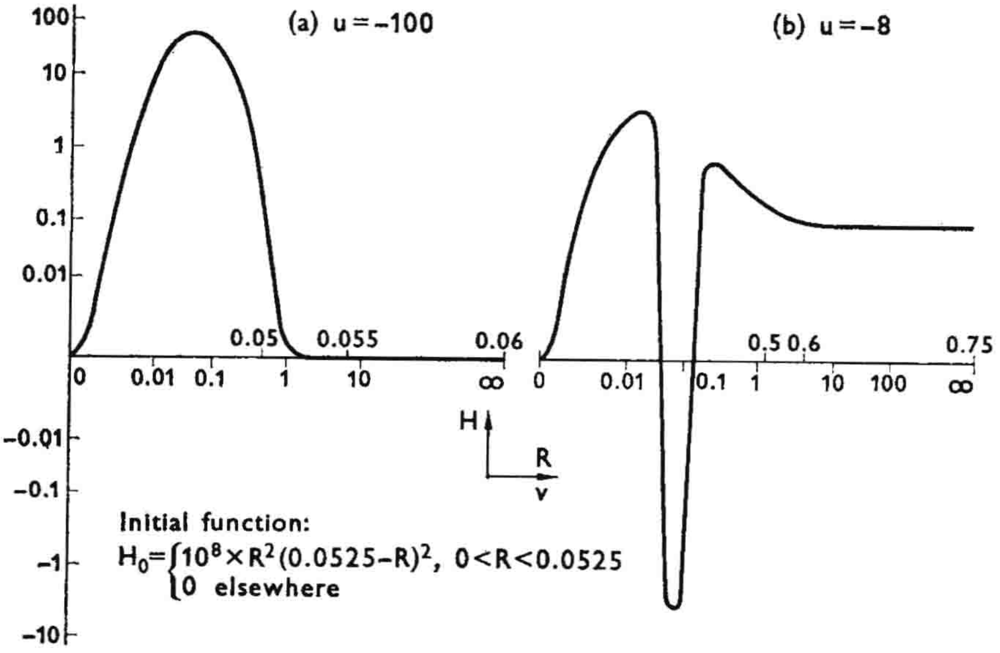


Figure 5a, b.

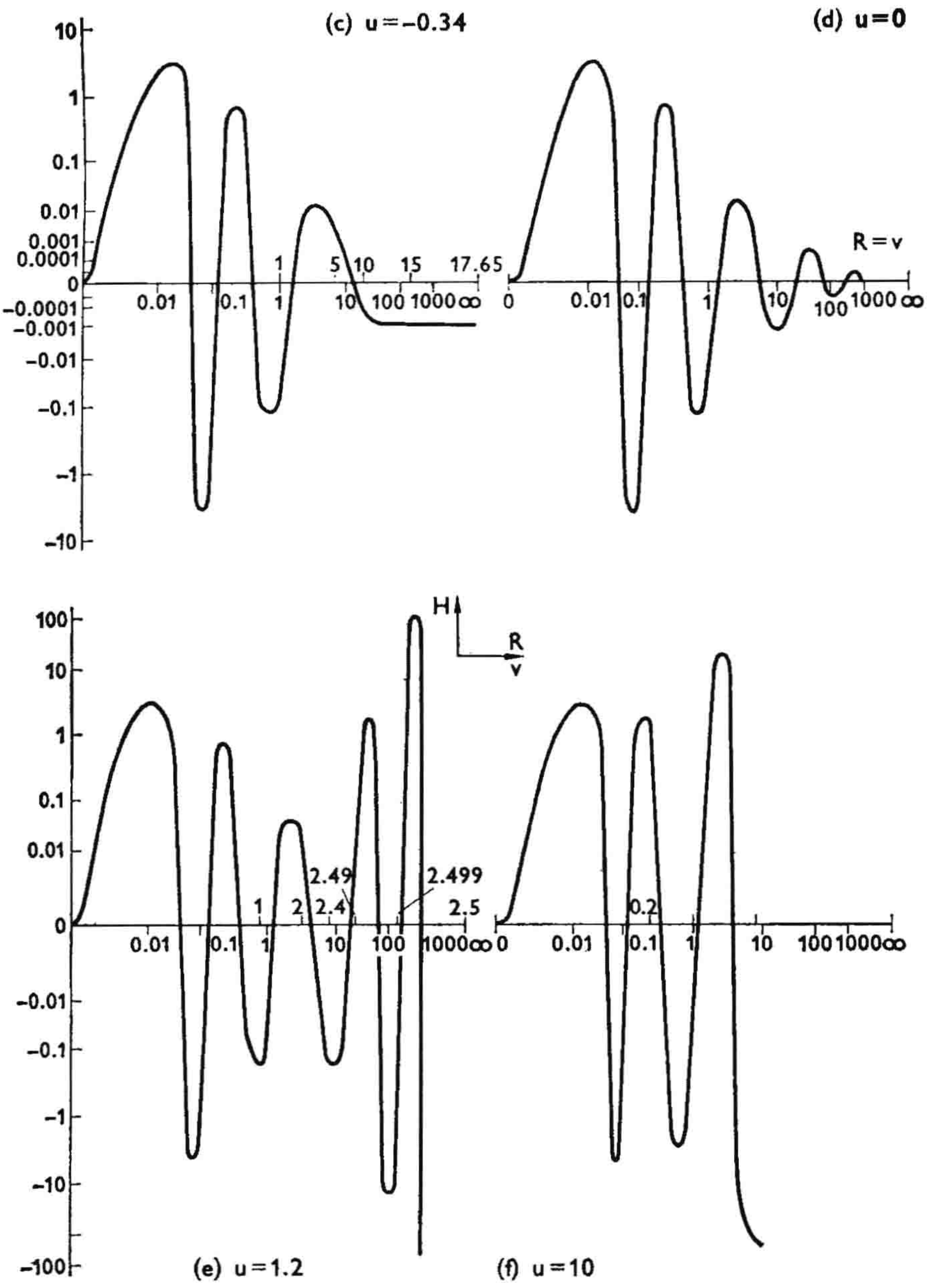


Figure 5c, d, e, f.

infinite number of oscillations, the amplitude of these tending to zero as $u \to 0$. For $u > 0$ however the solution curve waggles, showing no sign of tending to any limit as v rises.

The numerical procedure is a stable one, alterations of the step lengths in either direction producing no significant deviation from the graphs shown. The numerical evidence therefore points to a singularity in the test field along $r = r_-$ despite the fact that the solution appears to have the limiting value zero as $u \to 0$ and $v \to \infty$ along any curve for which $u \leqslant 0$.

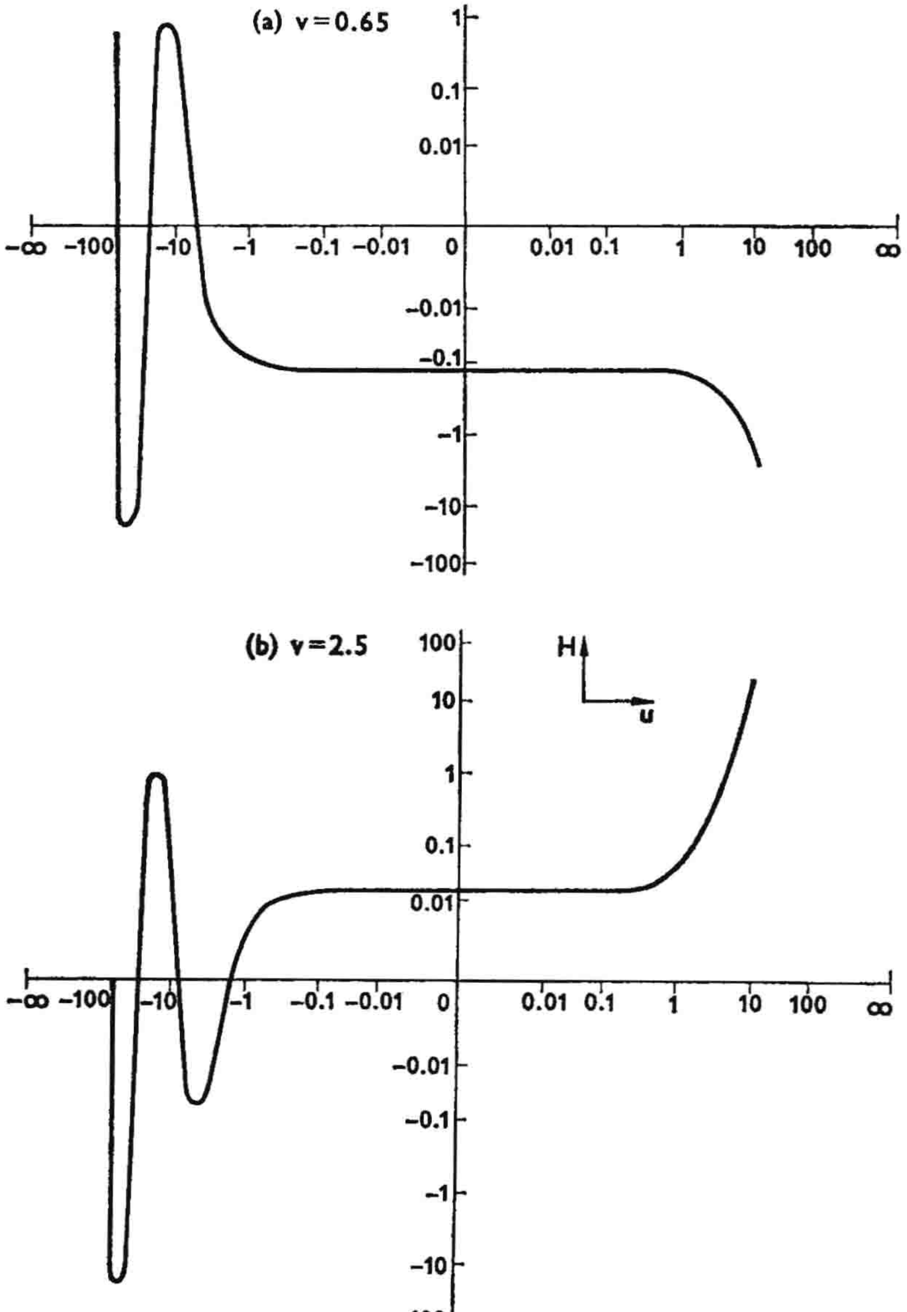


Figure 6a, b.

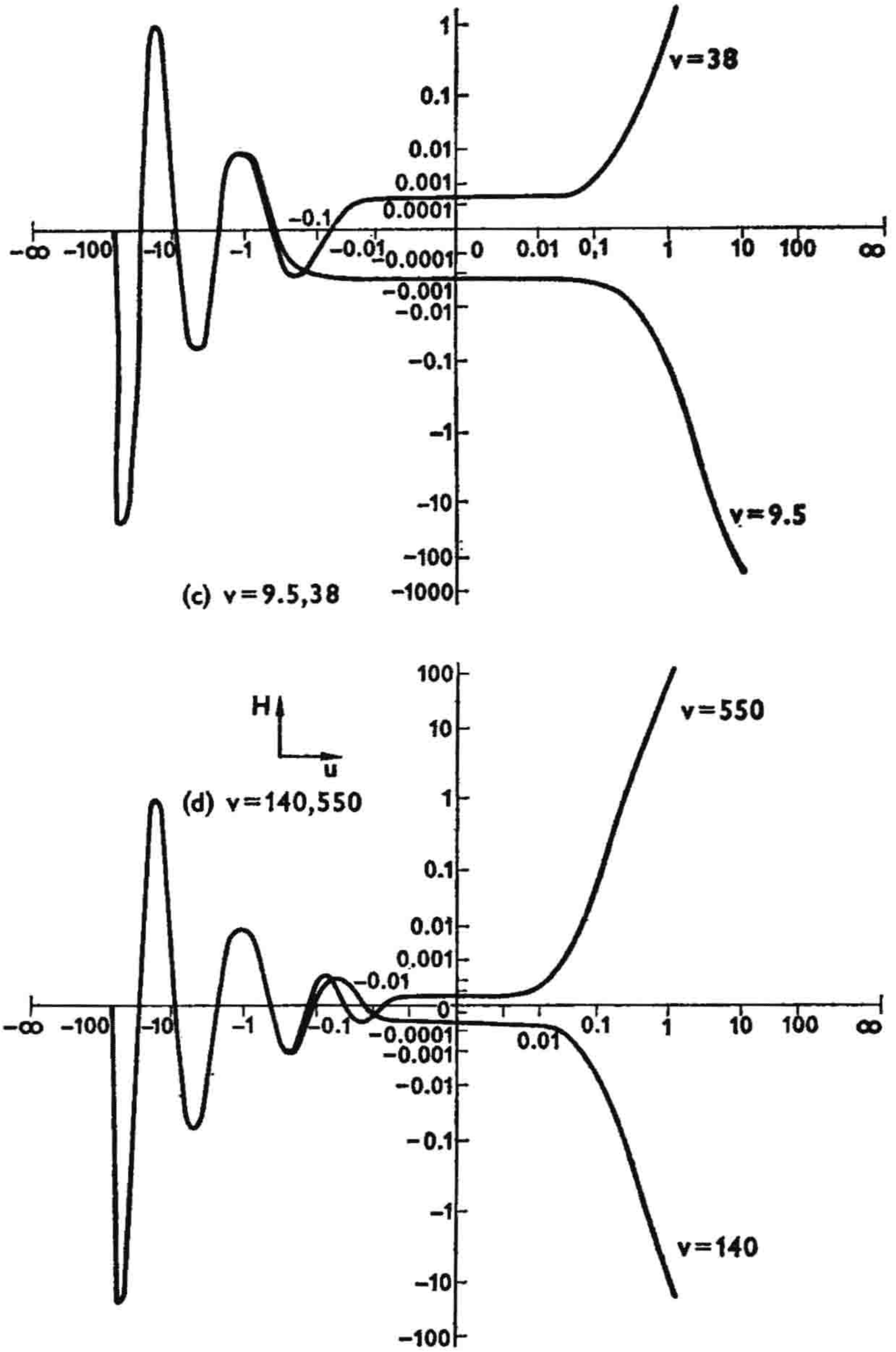


Figure 6c, d.

It is to be expected from analytical considerations of the characteristic initial value problem that the specification of unrestricted null characteristic data will in general produce a shock-wave up the initial hypersurface along which the field equations are violated, i.e. a sheet of charge. This will be so unless subsidiary conditions are imposed (namely the vanishing of certain integrals along the generators of the initial characteristic surface) in order to ensure that the sheet of charge is absent (the shock-wave cannot be directly ascertained from the numerical solution; the apparent initial severity of the graphs in Fig. 6 is partly due to the scale employed).

It is not to be expected that the presence of such a shock-wave should affect our conclusions, since it is only the effect of back-scattering from the wave which enters the region $u > 0$ in which we are mainly interested. However, to make sure of this we ran programs with suitable alternative initial data. In the first instance the effect of such a shock-wave can be substantially lessened and possibly even removed by balancing the initial 'hump' with another of equal size and opposite sign placed further up the hypersurface. The shock-wave can presumably be further reduced by separating two such positive humps with a larger negative hump so that a balance is again achieved (in the first case $\int_{v=v_0}^{v=\infty} H_0\,dR = 0$; in the second $\int_{v=v_0}^{v=\infty} H_0\,dR = \int_{v=v_0}^{v=\infty} [\partial H/\partial u]_{u=u_0}\,dR = 0$). It can also be removed from $u = u_0$ altogether by specifying non-zero data only along $v = v_0$ $(u < 0)$. The shock-wave then resides further along the hypersurface $v = v_0$. There is no reason to believe that the presence of this shock-wave should affect the field at $v = \infty$, $0 \leqslant u < \infty$.

Neither in these cases nor in any others considered has there been any major qualitative difference from the results displayed here. Each set of initial data produces a markedly stable solution while $u \leqslant 0$ followed by immediate signs of instability on crossing into $u > 0$. It therefore seems likely that the field singularity is a property of the partial differential equation (18) rather than any particular initial function and will consequently occur for generic cases of the latter.

It should be pointed out that the situation we consider of placing a *test* solution of Maxwell's equations on a background space-time which satisfies the Einstein–Maxwell equations is not equivalent to considering a small (electromagnetic) perturbation of the Einstein–Maxwell equations. This is because cross-terms arise in the contribution of the Maxwell field to the background space-time via the energy-momentum tensor, producing changes in the background metric of the same order as that of the initial perturbation.† A full treatment of such a perturbation would lead to far more complex equations than those treated here, but we do not believe that the results would be substantially affected. A perturbation analysis cannot in any case give a definitive answer to the problem we consider, since the non-linear effects would ultimately have to be brought in.

From this connection of inferences we conclude that the projected journey of our hypothetical observer through $r = r_-$ looks liable to prove a dangerous undertaking, for the likelihood is that the Cauchy horizon is unstable and that unbounded curvatures will be met in its immediate neighbourhood as soon as the coupling effect of the field back on the space-time is taken into account.

† We are grateful to Dr. R. Wald for pointing this out to us.

References

Boyer, R. H. and Lindquist, R. W. (1967). *Journal of Mathematical Physics*, **8**, 265.
Carter, B. (1968). *Physical Review*, **174**, 1559.
Carter, B. (1971). *Physical Review Letters*, **26**, 331.
Hawking, S. W. (1966). *Proceedings of the Royal Society*, *A*, **294**, 511; (1967) **300**, 187.
Hawking, S. W. (1972). *Communications in Mathematical Physics*, **25**, 152.
Hawking, S. W. and Penrose, R. (1970). *Proceedings of the Royal Society*, *A*, **314**, 529.
Israel, W. (1967). *Physical Review*, **164**, 1776.
Israel, W. (1968). *Communications in Mathematical Physics*, **8**, 245.
Kerr, R. P. (1963). *Physical Review Letters*, **11**, 237.
Newman, E. T., Couch, E., Chinnapared, K., Exton, A., Prakash, A. and Torrence, R. (1965). *Journal of Mathematical Physics*, **6**, 918.
Newman, E. T. and Penrose, R. (1962). *Journal of Mathematical Physics*, **3**, 566.
Newman, E. T. and Penrose, R. (1968). *Proceedings of the Royal Society*, *A*, **305**, 175.
Penrose, R. (1967). 'An Analysis of the Structure of Space-Time' (Adams Prize Essay, Cambridge). Princeton.
Penrose, R. (1968). 'Structure of Space-Time', in *Battelle Recontres* (Ed. C. M. De Witt and J. A. Wheeler). Benjamin, New York.
Penrose, R. (1969a). *Rivista del nuovo cimento*, Serie 1, **1**, Numero Speciale, 252.
Penrose, R. (1969b). *Contemporary Physics*, **1**, 545 (International Atomic Energy Agency, Vienna.
Smith, G. D. (1969). *Numerical Solution of Partial Differential Equations*. Oxford University Press.

58

A SPACE-TIME CALCULUS BASED ON PAIRS OF NULL DIRECTIONS

J. Math. Phys. **14(7)** (1973), 874–81. With R. Geroch and A. Held

A streamlined version of the spin-coefficient formalism of Chapter 15.

A space-time calculus based on pairs of null directions*

R. Geroch

Department of Physics, University of Chicago, Chicago, Illinois 60637

A. Held

Department of Physics, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

R. Penrose

Department of Mathematics, Birkbeck College, London, England

(Received 21 July 1972)

A formalism is presented for the treatment of space–times, which is intermediate between a fully covariant approach and the spin-coefficient method of Newman and Penrose. With the present formalism, a pair of null directions only, rather than an entire null tetrad, is singled out at each point. The concept of a spin- and boost-weighted quantity is defined, the formalism operating entirely with such quantities. This entails the introduction of modified differentiation operators, one of which represents a natural extension of the definition of the operator $\eth$ which had been introduced earlier by Newman and Penrose. For suitable problems, the present formalism should lead to considerable simplifications over that achieved by the standard spin-coefficient method.

1. INTRODUCTION

Certain types of calculation in general relativity are conveniently carried our using a tetrad formalism[1]: four vector fields are introduced which are linearly independent at each point. By taking components with respect to these vector fields any tensor field or tensor equation may be replaced by a system of scalar fields or scalar equations. In fact, ordinary calculations using tensor components according to some coordinate system are really a special case, since here we may use the gradients of the four scalar fields x^0, x^1, x^2, x^3 as the (covariant) tetrad vectors. (We should be careful to distinguish an actual tensor from its set of components in some coordinate or tetrad system.) For definiteness we shall adopt the "abstract index" conventions[2] according to which $\chi_{ab}{}^{c}{}_{d}$ denotes a tensor whereas $\chi_{\mathbf{ab}}{}^{\mathbf{c}}{}_{\mathbf{d}}$ dedenotes its set of components in a tetrad or coordinate system.[3] For example, a single vector field ξ_a may be described in terms of four scalar fields $\xi_0, \xi_1, \xi_2, \xi_3$, where $\xi_{\mathbf{a}} = \xi_a\delta^a_{\mathbf{a}}$, the tetrad system being denoted by $\delta^a_0, \delta^a_1, \delta^a_2, \delta^a_3$. Similarly, a tensor equation such as $\nabla_{(a}\xi_{b)} = 0$, for example, can be replaced, using this tetrad system, by the ten separate scalar equations

$$\nabla_{(\mathbf{a}}\xi_{\mathbf{b})} = \Gamma_{(\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b})}\xi_{\mathbf{c}}, \tag{1.1}$$

where $\nabla_0, \nabla_1, \nabla_2, \nabla_3$ are the tetrad components ("intrinsic derivatives") $\nabla_{\mathbf{a}} = \delta^a_{\mathbf{a}}\nabla_a$ of the covariant derivative operator ∇_a and where

$$\Gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}} = -\delta^a_{\mathbf{a}}\delta^b_{\mathbf{b}}\nabla_b\delta^{\mathbf{c}}_a. \tag{1.2}$$

Here, the *dual* tetrad $\delta^{\mathbf{a}}_a = \delta^0_a, \delta^1_a, \delta^2_a, \delta^3_a$, of $\delta^a_{\mathbf{a}}$, is being employed (defined by $\delta^{\mathbf{a}}_a\delta^a_{\mathbf{b}} = \delta^{\mathbf{a}}_{\mathbf{b}}$, that is, $\delta^{\mathbf{a}}_a\delta^a_{\mathbf{b}}$ is unity if $\mathbf{a} = \mathbf{b}$ and zero otherwise). For a general tetrad there will be *64* algebraically unrelated scalar fields $\Gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}}$. In the case of tetrads arising from coordinate systems we have the symmetry $\Gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}} = \Gamma_{\mathbf{b}}{}^{\mathbf{c}}{}_{\mathbf{a}}$ (Christoffel symbol symmetry), and we have just *40* algebraically unrelated scalar fields $\Gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}}$. On the other hand, if the tetrads are normalized according to some scheme:

$$\delta^a_{\mathbf{a}}\delta^b_{\mathbf{b}}g_{ab} = g_{\mathbf{ab}}, \tag{1.3}$$

where the $g_{\mathbf{ab}}$ are constants, being the elements of some numerical symmetric matrix [e.g., diag(1, − 1, − 1, − 1) for an orthonormal tetrad in a locally Minkowskian space], then we have the skew-symmetry

$$g_{\mathbf{cd}}\Gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}} = -g_{\mathbf{ca}}\Gamma_{\mathbf{d}}{}^{\mathbf{c}}{}_{\mathbf{b}}. \tag{1.4}$$

Thus we have only *24* independent scalar fields. These are called *Ricci rotation coefficients* (sometimes written $\gamma_{\mathbf{a}}{}^{\mathbf{c}}{}_{\mathbf{b}}$).

Since tetrad equations such as (1.1) are scalar equations, theorems on the existence and uniqueness of differential equations can be used directly. With a normalized tetrad, we have only 24 Γ's so we may expect some simplification over the use of coordinate components. Also, if one wishes to solve the gravitational equations explicitly, it is sometimes convenient to introduce normalized tetrads. Occasionally the equations become more transparent when written in terms of normalized tetrads. This may be the case when one or more of the vectors is *preferred;* for example, when a tetrad vector is chosen to be a Killing vector or, say, a multiple principal null vector of the Weyl tensor. But often equations become more difficult to interpret when written out explicitly in terms of tetrads. In the completely covariant formalism, every expression has a definite interpretation (in principle, at least) in terms of the geometry of the space–time, while in an explicit tetrad formalism, many of the equations are simply expressing relations between (perhaps arbitrarily chosen) tetrad vectors.

The spin-coefficient formalism[4] (referred to henceforth as the NP formalism) may be regarded, in the present context, as arising from a special case of a normalized tetrad formalism. As a consequence, there are 24 independent rotation coefficients; but it turns out that these may be combined in pairs to give just *12* independent *complex* scalar fields called *spin coefficients*. Real equations combine together in a natural way and it becomes a feasible proposition to assign a different letter to each spin coefficient and write out all equations explicitly in terms of components.

A common feature, then of the spin coefficient and tetrad formalisms is that they both involve a choice of a complete tetrad at each point. Often this choice has to be made arbitrarily to some extent since the geometry of the situation may not define that much structure at each point in a natural way. However, in many problems there may be some smaller amount of structure naturally singled out at each point. In particular, it is frequently the case that, say, one or two vectors or directions are naturally defined at each point, but not four vectors. In radiation problems, or when one is dealing with spacelike (or timelike) 2-surfaces, it is often the case that two (or one) *null directions* are singled out at each point by the conditions of the problem. The formalism that we set forth in this paper is designed specifically to deal with such situations.

We suppose that two of our future-pointing null directions are assigned at each point of the space–time. We can then choose, as two of our tetrad vectors, two null vectors l^a and n^a which point in these two directions. We can choose a normalization $l^a n_a = 1$ and take our remaining tetrad vectors to be unit spacelike vectors X^a and Y^a orthogonal to each of l^a and n^a and to each other. However, there is a two-dimensional "gauge" freedom remaining at each point, namely the 2-parameter subgroup of the Lorentz group at each point which preserves these two null directions. This group is generated by the boosts

$$l^a \to r l^a, \qquad n^a \to r^{-1} n^a \tag{1.5}$$

and the spatial rotations

$$m^a \to e^{i\theta} m^a, \tag{1.6}$$

where the complex vector m^a is defined by $m^a = 2^{-1/2}(X^a + iY^a)$. It is convenient to combine r and θ together in the form of the complex number λ where $\lambda^2 = r e^{i\theta}$. Then the "gauge group" at each point is seen to be the multiplicative group of complex numbers λ. Our formalism will deal with scalars associated with such a tetrad $(l^a, m^a, \bar{m}^a, n^a)$ where the scalars undergo transformations[5]

$$\eta \to \lambda^p \bar{\lambda}^q \eta \tag{1.7}$$

whenever the tetrad is changed according to (1.5) and (1.6). Such a scalar will be called a *spin- and boost-weighted scalar of type* $\{p, q\}$ (or simply a scalar of type $\{p, q\}$). The *spin-weight* is $\frac{1}{2}(p - q)$ and the *boost weight* is $\frac{1}{2}(p + q)$.

The formalism may be though of as a compromise between the fully covariant formalism and the spin coefficient formalism. It exhibits, to a certain extent, the advantages (and some disadvantages) of each. The use of indices is avoided; equations frequently have direct geometrical interpretations. Since the space–times considered in practice often have one or two null directions singled out, it is hoped that the modification of the spin-coefficient formalism considered here will serve as a useful computational tool.

2. THE FORMALISM

Let o^A and ι^A be a pair of spinor fields on the space–time normalized according to

$$o_A \iota^A = 1. \tag{2.1}$$

Such a pair of spinor fields is called a *dyad* or *spin frame*. It is well known[4] that any dyad defines a unique null tetrad $(l^a, m^a, \bar{m}^a, n^a)$ at each point and, conversely, that any null tetrad defines a dyad uniquely up to sign. The relationship is as follows[3]:

$$l^a = o^A \bar{o}^{A'}, \qquad m^a = o^A \bar{\iota}^{A}, \qquad \bar{m}^a = \iota^A \bar{o}^{A'}, \qquad n^a = \iota^A \bar{\iota}^{A'}. \tag{2.2}$$

This is clearly unaffected by the replacement $(o^A, \iota^A) \to (-o^A, -\iota^A)$. [A choice of null tetrad is equivalent to a choice of orthonormal tetrad. Canonically, we can set $T^a = 2^{-1/2}(l^a + n^a)$, $X^a = 2^{-1/2}(m^a + \bar{m}^a)$, $Y^a = -i2^{-1/2}(m^a - \bar{m}^a)$, and $Z^a = 2^{-1/2}(l^a - n^a)$ to define an orthonormal tetrad.] The twelve spin coefficients and their relation to the Ricci rotation coeefficients for the null tetrad are as follows[4]:

$$\begin{aligned}
\kappa &= o^A \bar{o}^{A'} o^B \nabla_{AA'} o_B = m^b l^a \nabla_a l_b, \\
\sigma &= o^A \bar{\iota}^{A'} o^B \nabla_{AA'} o_B = m^b m^a \nabla_a l_b, \\
\rho &= \iota^A \bar{o}^{A'} o^B \nabla_{AA'} o_B = m^b \bar{m}^a \nabla_a l_b, \\
\tau &= \iota^A \bar{\iota}^{A'} o^B \nabla_{AA'} o_B = m^b n^a \nabla_a l_b, \\
\kappa' &= -\iota^A \bar{\iota}^{A'} \iota^B \nabla_{AA'} \iota_B = \bar{m}^b n^a \nabla_a n_b, \\
\sigma' &= -\iota^A \bar{o}^{A'} \iota^B \nabla_{AA'} \iota_B = \bar{m}^b \bar{m}^a \nabla_a n_b \\
\rho' &= -o^A \bar{\iota}^{A'} \iota^B \nabla_{AA'} \iota_B = \bar{m}^b m^a \nabla_a n_b, \\
\tau' &= -o^A \bar{o}^{A'} \iota^B \nabla_{AA'} \iota_B = \bar{m}^b l^a \nabla_a n_b
\end{aligned} \tag{2.3}$$

and

$$\begin{aligned}
\beta &= o^A \bar{\iota}^{A'} \iota^B \nabla_{AA'} o_B = \tfrac{1}{2}(n^b m^a \nabla_a l_b - \bar{m}^b m^a \nabla_a m_b), \\
\beta' &= -\iota^A \bar{o}^{A'} o^B \nabla_{AA'} \iota_B = \tfrac{1}{2}(l^b \bar{m}^a \nabla_a n_b - m^b \bar{m}^a \nabla_a \bar{m}_b), \\
\epsilon &= o^A \bar{o}^{A'} \iota^B \nabla_{AA'} o_B = \tfrac{1}{2}(n^b l^a \nabla_a l_b - \bar{m}^b l^a \nabla_a m_b), \\
\epsilon' &= -\iota^A \bar{\iota}^{A'} o^B \nabla_{AA'} \iota_B = \tfrac{1}{2}(l^b n^a \nabla_a n_b - m^b n^a \nabla_a \bar{m}_b).
\end{aligned} \tag{2.4}$$

The above notation differs from that originally given by Newman and Penrose in that we have used only six different Greek letters rather than twelve. The use of primed letters brings out the close relation between six of the spin coefficients and the remaining six. The other six Greek symbols of the NP formalism are related to ours by

$$\begin{aligned}
\nu &= -\kappa', \qquad \lambda = -\sigma', \qquad \mu = -\rho', \\
\pi &= -\tau', \qquad \alpha = -\beta', \qquad \gamma = -\epsilon'.
\end{aligned} \tag{2.5}$$

We shall make use of the prime systematically here to denote the operation of effecting the replacement:

$$o^A \to i\iota^A, \; \iota^A \to i o^A, \qquad \bar{o}^{A'} \to -i\bar{\iota}^{A'}, \; \bar{\iota}^{A'} \to -i\bar{o}^{A'}, \tag{2.6}$$

so that

$$(l^a)' = n^a, \qquad (m^a)' = \bar{m}^a, \qquad (\bar{m}^a)' = m^a, \qquad (n^a)' = l^a. \tag{2.7}$$

This preserves the normalization (2.1) and the relationship between a quantity and its complex conjugate. Since the bar and prime operations commute, one can write $\bar{\eta}'$ without ambiguity. Furthermore the prime operation is involutory up to sign:

$$(\eta')' = (-1)^{p+q}\eta. \tag{2.7a}$$

(For all quantities explicitly defined in this paper, $p + q$ is in fact even, so this sign will play no role here.) This use of the prime not only halves the number of Greek letters needed, but also effectively halves the number of equations.[6]

The role of the (vector) covariant derivative operator ∇_a is taken over in the NP formalism by four scalar operators:

$$\begin{aligned}
D &= o^A \bar{o}^{A'} \nabla_{AA'} = l^a \nabla_a, \qquad D' = \iota^A \bar{\iota}^{A'} \nabla_{AA'} = n^a \nabla_a, \\
\delta &= o^A \bar{\iota}^{A'} \nabla_{AA'} = m^a \nabla_a, \qquad \bar{\delta} = \delta' = \iota^A \bar{o}^{A'} \nabla_{AA'} = \bar{m}^a \nabla_a.
\end{aligned} \tag{2.8}$$

Tensor or spinor equations on the space–time can now be written out explicitly in terms of the components with respect to the null tetrad or the dyad, and the spin coefficients (2.3), (2,4), and the derivative operators (2.8).

So far this is just standard NP formalism. The innovation that we shall introduce here, however, is to work entirely with spin- and boost-weighted quantities, since these are the quantities which are apropriately "covariant" under the group of tetrad transformations which leave the two null directions of l^a and n^a invariant. Now the most general spin transformation which leaves these two null directions invariant [preserving the normalization (2.1)] is

$$o^A \to \lambda o^A, \qquad \iota^A \to \lambda^{-1}\iota^A, \tag{2.9}$$

where λ is an arbitrary (nowehere vanishing) complex scalar field. In terms of the null tetrad, the transformation (2.9) takes the form

$$\begin{aligned} &l^a \to \lambda\bar{\lambda} l^a, \qquad m^a \to \lambda\bar{\lambda}^{-1} m^a, \\ &\bar{m}^a \to \lambda^{-1}\bar{\lambda}\bar{m}^a, \qquad n^a \to \lambda^{-1}\bar{\lambda}^{-1} n^a. \end{aligned} \tag{2.10}$$

These are just the transformations (1.5) and (1.6) considered earlier.

Recall that a scalar η of type $\{p, q\}$ was to be a scalar quantity which transformed according to

$$\eta \to \lambda^p \bar{\lambda}^q \eta, \tag{2.11}$$

whenever the dyad (o^A, ι^A) transformed according to (2.9) or, equivalently, whenever the null tetrad $(l^a, m^a, \bar{m}^a, n^a)$ transformed according to (2.10). Strictly speaking, we should think of η as a *function* which assigns a complex scalar field $\eta(o^A, \iota^A)$ to each ordered normalized pair of spinor fields o^A, ι^A, for which the null directions defined by o^A and by ι^A are the given pair of null directions. Alternatively, we can think of η as a function of the null tetrad. In any case, to be a proper spin- and boost-weighted quantity, the function η must be of a very special type, namely one which satisfies (2.11) when the dyad or null tetrad is changed according to (2.9) or (2.10), i.e. $\eta(\lambda o^A, \lambda^{-1}\iota^A) = \lambda^p\bar{\lambda}^q\eta(o^A, \iota^A)$. (Note that we may regard o^A and ι^A, themselves as *spinors* of type $\{1, 0\}$ and $\{-1, 0\}$, respectively, and $l^a, m^a, \bar{m}^a, n^a$ as *vectors* of type $\{1, 1\}$, $\{1, -1\}$, $\{-1, 1\}$ and $\{-1, -1\}$, respectively.)

The spin coefficients may now be divided into two classes according to whether or not they are proper spin- and boost-weighted quantities. In fact, the spin coefficients in the list (2.3) are all such quantities whereas those in the list (2.4) are not. Let us illustrate this with two examples:

$$\sigma \to (\lambda o^A)(\bar{\lambda}^{-1}\bar{\iota}^{A'})(\lambda o^B)\nabla_{AA'}(\lambda o_B) = \lambda^3\bar{\lambda}^{-1}\sigma, \tag{2.12a}$$

but

$$\begin{aligned} \beta \to (\lambda o^A)&(\bar{\lambda}^{-1}\bar{\iota}^{A'})(\lambda^{-1}\iota^B)\nabla_{AA'}(\lambda o_B) \\ &= \lambda\bar{\lambda}^{-1}\beta + \bar{\lambda}^{-1}o^A\bar{\iota}^{A'}\nabla_{AA'}\lambda. \end{aligned} \tag{2.12b}$$

The types of the spin- and boost-weighted quantities (2.3) are as follows:

$$\begin{aligned} &\kappa{:}\{3, 1\}, \sigma{:}\{3, -1\}, \rho{:}\{1, 1\}, \tau{:}\{1, -1\}, \\ &\kappa'{:}\{-3, -1\}, \sigma'{:}\{-3, 1\}, \rho'{:}\{-1, -1\}, \tau'{:}\{-1, 1\}. \end{aligned} \tag{2.13}$$

With any spinor field or tensor field on the space–time there is associated a collection of spin- and boost-weighted scalars of various types $\{p, q\}$ which together define the spinor or tensor. These are obtained from the spinor by transvecting with the various combinations of o^A, ι^A, $\bar{o}^{A'}$, and $\bar{\iota}^{A'}$, or the tensor with the various combinations of $l^a, m^a, \bar{m}^a$, and n^a. Any tensor field, may, in fact, be interpreted as a spinor field if desired; but we get precisely the same set of scalars whichever way we do it because of the definition (2.2) of the null tetrad in terms of the dyad.

Evidently the product of a scalar of type $\{p, q\}$ with a scalar of type $\{u, v\}$ is a scalar of type $\{p + u, q + v\}$. On the other hand, sums are allowed only when the summands have the same type. The type of the sum is the same as that of each summand.

We next wish to introduce derivative operators into the formalism. Unfortunately the operators (2.8) of the NP formalism are not suitable for this purpose, for, when applied to a scalar of type $\{p, q\}$, with p, q not both zero, they do not in general produce a spin- and boost-weighted scalar. We, therefore, modify the derivative operators (2.8) by the inclusion of further terms involving the spin coefficients in the list (2.4). For η of type $\{p, q\}$ we define[7,8]

$$\begin{aligned} &\text{þ}\eta = (D - p\epsilon - q\bar{\epsilon})\eta, \qquad \text{þ}'\eta = (D' + p\epsilon' + q\bar{\epsilon}')\eta, \\ &\eth\eta = (\delta - p\beta + q\bar{\beta}')\eta, \qquad \eth'\eta = (\delta' + p\beta' - q\bar{\beta})\eta. \end{aligned} \tag{2.14}$$

These combination have been so chosen that, under (2.9), the terms involving derivatives of λ cancel exactly. It should be noted that these operators are *derivations*, that is they are linear, and when operating on products, they satisfy the Leibniz rule. The operator $\eth$ as defined here reduces, in the appropriate circumstances, to that of Newman and Penrose.[9,10] The spin weights of the operators (2.14) are as follows:

$$\begin{aligned} &\text{þ}{:}\{1, 1\}, \qquad \eth{:}\{1, -1\}, \\ &\text{þ}'{:}\{-1, -1\}, \qquad \eth'{:}\{-1, 1\}. \end{aligned} \tag{2.15}$$

(To say that a differential operator has type $\{p, q\}$ is to say that when acting on a scalar of type $\{u, v\}$ it produces a scalar of type $\{p + u, q + v\}$.)

Alternatively the operators may be defined in terms of the type $\{0, 0\}$ operator (acting on a quantity of type $\{p, q\} = \{r + s, r - s\}$)

$$\begin{aligned} \Theta_{AA'} &= \nabla_{AA'} - p\iota^B\nabla_{AA'}o_B - q\bar{\iota}^{B'}\nabla_{AA'}\bar{o}_{B'} \\ &= \nabla_a - rn^b\nabla_a l_b + s\bar{m}^b\nabla_a m_b \end{aligned} \tag{2.14a}$$

by the equation

$$\Theta_a = l_a\text{þ}' + n_a\text{þ} - m_a\eth' - \bar{m}_a\eth. \tag{2.14b}$$

In Eq. (2.14a), s and r are the spin and boost weights, respectively [cf. following 1.7)]. The original definitions (2.14) may be recovered by transvecting (2.14b) with l^a, n^a, m^a and $\bar{m}^a$.

Since we wish to operate only with quantities with a well-defined $\{p, q\}$, we are not able to use the spin coefficients in the list (2.4) directly. Instead, the role of these spin coefficients is to be found within the definitions of the operators (2.14).

The basic quantities with which we shall work are the eight spin coefficients $\kappa, \sigma, \rho, \tau; \kappa', \sigma', \rho', \tau'$ and the four differential operators $\text{þ}, \eth, \text{þ}', \eth'$; In addition, there is the operation of complex conjugation. We may also consider the prime as effectively an allowable operation on the system. With the introduction of various tensor or spinor fields, such as the electromagnetic field tensor or Riemann tensor, we shall be able to combine the above elements with the tetrad components of the tensor fields and the dyad components of the spinor fields to obtain a self-contained calculus.

The effect of the derivative operators (2.14) is shown in Fig. 1. We associate with a scalar η of type $\{p,q\}$, the point with coordinates $\{p,q\}$ in the plane. Each of the derivative operators (2.14) has a characteristic effect on the type, which can be represented as a displacement in this diagram. Note that when two elements are multiplied together this corresponds to a vector sum in the diagram. If two elements are to be added together, then they must be represented by the same point in the diagram. The operation of complex conjugation is represented by a reflection in the line $p=q$, since the complex conjugate of an element of type $\{p,q\}$ is an element of type[18] $\{\bar{q},\bar{p}\}$. We define, in fact,

$$\bar{\text{þ}}=\text{þ},\qquad \bar{\text{þ}}'=\text{þ}',\qquad \bar{\eth}=\eth',\qquad \bar{\eth}'=\eth. \qquad (2.16)$$

Then the operation of complex conjugation will satisfy

$$\overline{\text{þ}\eta}=\bar{\text{þ}}\bar{\eta},\qquad \overline{\eth\eta}=\bar{\eth}\bar{\eta}. \qquad (2.17)$$

Finally, the prime operation is represented in the diagram by a reflection in the origin, since if we prime an element of type $\{p,q\}$ we get an element of type $\{-p,-q\}$. The prime will commute with addition, multiplication, and complex conjugation [but note (2.7a)]. Furthermore, we have

$$(\text{þ}\eta)'=\text{þ}'\eta',\qquad (\text{þ}'\eta)'=\text{þ}\eta',\qquad (\eth\eta)'=\eth'\eta',\qquad (\eth'\eta)'=\eth\eta'. \qquad (2.18)$$

The various spin-coefficient formulae will now be given explicitly in the form allowable within our present formalism. As an example, consider Eq. (4.2k) of Ref. 4:

$$\delta\rho-\delta'\sigma=\rho(\beta-\bar{\beta}')+\sigma(\bar{\beta}+3\beta') + \tau(\rho-\bar{\rho})+\kappa(\bar{\rho}'-\rho')-\Psi_1+\Phi_{01}. \qquad (2.19)$$

Rewriting (2.19) as

$$(\delta-\beta+\bar{\beta}')\rho-(\delta'+3\beta'+\bar{\beta})\sigma = (\rho-\bar{\rho})\tau+(\bar{\rho}'-\rho')\kappa-\Psi_1+\Phi_{01} \qquad (2.20)$$

and noting that ρ and σ are types $\{1,1\}$ and $\{3,-1\}$, respectively, we see that the equation may be reexpressed, using (2.14) as

$$\eth\rho-\eth'\sigma=(\rho-\bar{\rho})\tau+(\bar{\rho}'-\rho')\kappa-\Psi_1+\Phi_{01}. \qquad (2.21)$$

Similarly, Eqs. (4.2a), (4.2b), (4.2c), (4.2p). and (4.2q) of Ref. 4 may be rewritten as

$$\text{þ}\rho-\eth'\kappa=\rho^2+\sigma\bar{\sigma}-\bar{\kappa}\tau-\tau'\kappa+\Phi_{00}, \qquad (2.22)$$

$$\text{þ}\sigma-\eth\kappa=\sigma(\rho+\bar{\rho})-\kappa(\tau+\bar{\tau}')+\Psi_0, \qquad (2.23)$$

$$\text{þ}\tau-\text{þ}'\kappa=\rho(\tau-\bar{\tau}')+\sigma(\bar{\tau}-\tau')+\Psi_1+\Phi_{01}, \qquad (2.24)$$

$$\eth\tau-\text{þ}'\sigma=-\rho'\sigma-\bar{\sigma}'\rho+\tau^2+\kappa\bar{\kappa}'+\Phi_{02}, \qquad (2.25)$$

and

$$\text{þ}'\rho-\eth'\tau=\rho\bar{\rho}'+c\sigma'-\tau\bar{\tau}-\kappa\kappa'-\Psi_2-2\Lambda. \qquad (2.26)$$

Applying the prime operation to each of these six equations, we obtain six more equations, these being equivalent to (4.2m), (4.2n), (4.2j), (4.2i), (4.2g), and (4.2h) of Ref. 4, respectively. We recall the definitions of the dyad components of the Weyl spinor Ψ_{ABCD} and the trace-free Ricci spinor

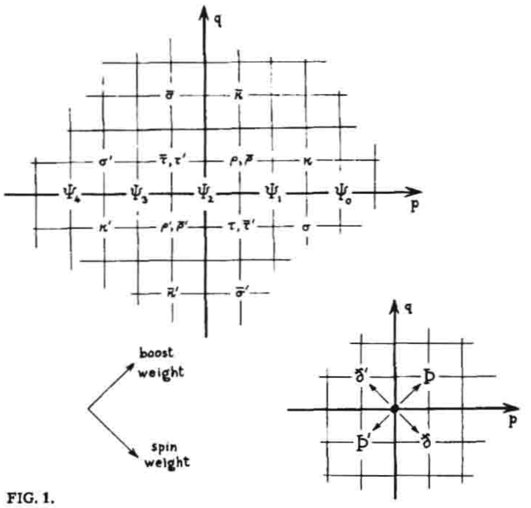


FIG. 1.

$\Phi_{ABC'D'}$:

$$\begin{aligned}
\Psi_0&=o^Ao^Bo^Co^D\Psi_{ABCD}=\Psi_4' && \{4,0\},\\
\Psi_1&=o^Ao^Bo^C\iota^D\Psi_{ABCD}=\Psi_3' && \{2,0\},\\
\Psi_2&=o^Ao^B\iota^C\iota^D\Psi_{ABCD}=\Psi_2' && \{0,0\},\\
\Psi_3&=o^A\iota^B\iota^C\iota^D\Psi_{ABCD}=\Psi_1' && \{-2,0\},\\
\Psi_4&=\iota^A\iota^B\iota^C\iota^D\Psi_{ABCD}=\Psi_0' && \{-4,0\};
\end{aligned} \qquad (2.27)$$

$$\begin{aligned}
\Phi_{00}&=o^Ao^B\bar{o}^{A'}\bar{o}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{00}=\Phi_{22}' && \{2,2\},\\
\Phi_{01}&=o^Ao^B\bar{o}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{10}=\Phi'_{21} && \{2,0\},\\
\Phi_{02}&=o^Ao^B\bar{\iota}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{20}=\Phi'_{20} && \{2,-2\},\\
\Phi_{10}&=o^A\iota^B\bar{o}^{A'}\bar{o}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{01}=\Phi'_{12} && \{0,2\},\\
\Phi_{11}&=o^A\iota^B\bar{o}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{11}=\Phi'_{11} && \{0,0\},\\
\Phi_{12}&=o^A\iota^B\bar{\iota}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{21}=\Phi'_{10} && \{0,-2\},\\
\Phi_{20}&=\iota^A\iota^B\bar{o}^{A'}\bar{o}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{02}=\Phi_{02}' && \{-2,+2\},\\
\Phi_{21}&=\iota^A\iota^B\bar{o}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{12}=\Phi_{01}' && \{-2,0\},\\
\Phi_{22}&=\iota^A\iota^B\bar{\iota}^{A'}\bar{\iota}^{B'}\Phi_{ABA'B'}=\bar{\Phi}_{22}=\Phi_{00}' && \{-2,-2\}.
\end{aligned} \qquad (2.28)$$

In each case, the type is given on the right. Furthermore, the scalar curvature is defined by

$$\Lambda=\bar{\Lambda}=\Lambda'=\tfrac{1}{24}R \qquad \{0,0\}. \qquad (2.29)$$

The list (2.21)–(2.26), together with the corresponding list of primed equations does not completely exhaust the NP equations (4.2) of Ref. 4. The remaining equations (4.2d), (4.2e), (4.2f), (4.2l), (4.2o), and (4.2r) in Ref. 4 refer to derivatives of spin coefficients which are not spin- and boost-weighted quantities. They cannot, therefore, be written explicitly in our present formalism as equations like (2.21). Instead, they play their role as part of the commutator equations for the differential operators þ, þ′, ð and ð′. These commutators, when applied to a spin- and boost-weighted scalar η of type $\{p,q\}$, are

$$(\text{Þ}\text{Þ}' - \text{Þ}'\text{Þ})\eta = [(\bar{\tau} - \tau')\eth + (\tau - \bar{\tau}')\eth' - p(\kappa\kappa' - \tau\tau' + \Psi_2 + \Phi_{11} - \Lambda) - q(\bar{\kappa}\bar{\kappa}' - \overline{\tau\tau}' + \bar{\Psi}_2 + \Phi_{11} - \Lambda)]\eta, \tag{2.30}$$

$$(\text{Þ}\eth - \eth\text{Þ})\eta = [\bar{\rho}\eth + \sigma\eth' - \bar{\tau}'\text{Þ} - \kappa\text{Þ}' - p(\rho'\kappa - \tau'\sigma + \Psi_1) - q(\bar{\sigma}'\bar{\kappa} - \overline{\rho\tau}' + \Phi_{01})]\eta, \tag{2.31}$$

$$(\eth\eth' - \eth'\eth)\eta = [(\bar{\rho}' - \rho')\text{Þ} + (\rho - \bar{\rho})\text{Þ}' + p(\rho\rho' - \sigma\sigma' + \Psi_2 - \Phi_{11} - \Lambda) - q(\overline{\rho\rho}' - \overline{\sigma\sigma}' + \Psi_2 - \Phi_{11} - \Lambda)]\eta, \tag{2.32}$$

together with the remaining commutator equations obtained by applying prime, complex conjugation, and both to (2.31). Note that the spin- and boost-weights of η enter explicitly on the right-hand side. We must be careful, when applying primes and bars to these equations, to remember that η', $\bar{\eta}$, and $\bar{\eta}'$ have types which are not quite those of η. Thus, under the prime, p becomes $-p$ and q becomes $-q$; under the bar, p becomes q and q becomes p; under both bar and prime p becomes $-q$ and q becomes $-p$.

The commutator equations are the one place where our present formalism yields more complicated formulas than the original NP formalism. This seems to be the price we pay for the very considerable formal simplification that we obtain for the other equations. But we must bear in mind that our commutators are actually combining information which comes from two different places in the NP formalism. There is, however, a gain as regards geometric content of the commutators with our present formulation. The extra terms which arise when p or q is nonzero may sometimes be interpreted as curvature quantities referring to submanifolds in the space-time. We shall see this explicitly for Eq. (2.32) in the next section.

The full Bianchi identities consist of[11]

$$\text{Þ}\Psi_1 - \eth'\Psi_0 - \text{Þ}\Phi_{01} + \eth\Phi_{00} = -\tau'\Psi_0 + 4\rho\Psi_1 - 3\kappa\Psi_2 + \bar{\tau}'\Phi_{00} - 2\bar{\rho}\Phi_{01} - 2\sigma\Phi_{10} + 2\kappa\Phi_{11} + \bar{\kappa}\Phi_{02}, \tag{2.33}$$

$$\text{Þ}\Psi_2 - \eth'\Psi_1 - \eth'\Phi_{01} + \text{Þ}'\Phi_{00} + 2\text{Þ}\Lambda = \sigma'\Psi_0 - 2\tau'\Psi_1 + 3\rho\Psi_2 - 2\kappa\Psi_3 + \bar{\rho}'\Phi_{00} - 2\bar{\tau}\Phi_{01} - 2\tau\Phi_{10} + 2\rho\Phi_{11} + \bar{\sigma}\Phi_{02}, \tag{2.34}$$

$$\text{Þ}\Psi_3 - \eth'\Psi_2 - \text{Þ}\Phi_{21} + \eth\Phi_{20} - 2\eth'\Lambda = 2\sigma'\Psi_1 - 3\tau'\Psi_2 + 2\rho\Psi_3 - \kappa\Psi_4 - 2\rho'\Phi_{10} + 2\tau'\Phi_{11} + \bar{\tau}'\Phi_{20} - 2\bar{\rho}\Phi_{21} + \bar{\kappa}\Phi_{22}, \tag{2.35}$$

$$\text{Þ}\Psi_4 - \eth'\Psi_3 - \eth'\Phi_{21} + \text{Þ}'\Phi_{20} = +3\sigma'\Psi_2 - 4\tau'\Psi_3 + \rho\Psi_4 - 2\kappa'\Phi_{10} + 2\sigma'\Phi_{11} + \bar{\rho}'\Phi_{20} - 2\bar{\tau}\Phi_{21} + \bar{\sigma}\Phi_{22} \tag{2.36}$$

together with their primed versions, the contracted Bianchi identities

$$\text{Þ}\Phi_{11} + \text{Þ}'\Phi_{00} - \eth\Phi_{10} - \eth'\Phi_{01} + 3\text{Þ}\Lambda = (\rho' + \bar{\rho}')\Phi_{00} + 2(\rho + \bar{\rho})\Phi_{11} - (\tau' + 2\bar{\tau})\Phi_{01} - (2\tau + \bar{\tau}')\Phi_{10} - \bar{\kappa}\Phi_{12} - \kappa\Phi_{21} + \sigma\Phi_{20} + \bar{\sigma}\Phi_{02} \tag{2.37}$$

$$\text{Þ}\Phi_{12} + \text{Þ}'\Phi_{01} - \eth\Phi_{11} - \eth'\Phi_{02} + 3\eth\Lambda = (\rho' + 2\bar{\rho}')\Phi_{01} + (2\rho + \bar{\rho})\Phi_{12} - (\tau' + \bar{\tau})\Phi_{02} - 2(\tau + \bar{\tau}')\Phi_{11} - \bar{\kappa}'\Phi_{00} - \kappa\Phi_{22} + \sigma\Phi_{21} + \bar{\sigma}'\Phi_{10} \tag{2.38}$$

and the primed versions of these equations. [More generally, equations similar to (2.37) and (2.38) express the equations of an arbitrary conserved symmetric two index tensor.]

The content of Einstein's vacuum equations is obtained by putting all the Φ's and Λ's equal to zero in (2.21)–(2.26) and in (2.30)–(2.32). The Bianchi identities (2.33)–(2.36) (with $\Phi = \Lambda = 0$) become considerably simpler in this case and closely resemble Maxwell's source-free equations. In the present formalism, Maxwell's source-free equations are

$$\text{Þ}\phi_1 - \eth'\phi_0 = -\tau'\phi_0 + 2\rho\phi_1 - \kappa\phi_2, \tag{2.39}$$

$$\text{Þ}\phi_2 - \eth'\phi_1 = \sigma'\phi_0 - 2\tau'\phi_1 + \rho\phi_2, \tag{2.40}$$

together with their primed versions. Here we have

$$\begin{aligned} \phi_0 &= o^A o^B \phi_{AB} = -\phi_2' \quad \{2,0\},\\ \phi_1 &= o^A \iota^B \phi_{AB} = -\phi_1' \quad \{0,0\},\\ \phi_2 &= \iota^A \iota^B \phi_{AB} = -\phi_0' \quad \{-2,0\}, \end{aligned} \tag{2.41}$$

where the symmetric spinor ϕ_{AB} is related to Maxwell's field tensor by

$$F_{ab} = \phi_{AB}\epsilon_{A'B'} + \epsilon_{AB}\bar{\phi}_{A'B'},$$

Finally, we remark on the existence of an additional symmetry possessed by the spin-coefficient formalism which was noticed some time ago by Sachs.[12] We denote the Sachs symmetry by an asterisk (*), defining

$$(o^A)^* = o^A, \quad (\iota^A)^* = \iota^A, \quad (\bar{o}^{A'})^* = \bar{\iota}^{A'}, \quad (\bar{\iota}^{A'})^* = -\bar{o}^{A'}; \tag{2.42}$$

so that

$$\begin{aligned} (l^a)^* &= m^a, \quad (m^a)^* = -l^a,\\ (\bar{m}^a)^* &= n^a, \quad (n^a)^* = -\bar{m}^a. \end{aligned} \tag{2.43}$$

This preserves the required normalizations and orthogonality relations). Clearly the Sachs symmetry operation does not commute with complex conjugation, since the relation between an object and its complex conjugate is destroyed. However, we do have

$$(n^*)^* = (-1)^q\eta, \quad (\eta')^* = (-1)^q(\eta^*)',$$

$$\text{and } \overline{\eta^*} = (-i)^{p+q}(\bar{\eta}')^*, \tag{2.44}$$

where η is the quantity of type $\{p, q\}$. Note for such an η, the quantity η^* is of type $\{p, -q\}$.

From (2.3) we get

$$\begin{aligned} &\kappa^* = \sigma, \quad \sigma^* = -\kappa, \quad \rho^* = \tau, \quad \tau^* = -\rho, \quad \kappa'^* = -\sigma',\\ &\quad \sigma'^* = \kappa', \quad \rho'^* = -\tau', \tau'^*, = \rho',\\ &\bar{\kappa}^* = -\bar{\sigma}', \quad \bar{\sigma}^* = -\bar{\kappa}', \quad \bar{\rho}^* = \tau', \quad \bar{\tau}^* = \bar{\rho}', \quad \bar{\kappa}'^* = \bar{\sigma},\\ &\quad \bar{\sigma}'^* = \bar{\kappa}, \quad \bar{\rho}'^* = -\bar{\tau}, \quad \bar{\tau}'^* = -\bar{\rho}, \end{aligned} \tag{2.45}$$

$$\text{Þ}^* = \eth, \quad \eth^* = -\text{Þ}, \quad \text{Þ}'^* = -\eth', \quad \eth'^* = \text{Þ}', \tag{2.46}$$

$$\Psi_0{}^* = \Psi_0, \quad \Psi_1{}^* = \Psi_1, \quad \Psi_2{}^* = \Psi_2,$$

$$\Psi_3{}^* = \Psi_3, \quad \Psi_4{}^* = \Psi_4,$$
$$\overline{\Psi}_0{}^* = \overline{\Psi}_4, \quad \overline{\Psi}_1{}^* = -\overline{\Psi}_3, \quad \overline{\Psi}_2{}^* = \overline{\Psi}_2,$$
$$\overline{\Psi}_3{}^* = -\overline{\Psi}_1, \quad \overline{\Psi}_4{}^* = \overline{\Psi}_0, \tag{2.47}$$

$$\Phi_{00}{}^* = \Phi_{02}, \quad \Phi_{01}{}^* = -\Phi_{01}, \quad \Phi_{02}{}^* = \Phi_{00}, \Phi_{10}{}^* = \Phi_{12},$$
$$\Phi_{11}{}^* = -\Phi_{11}, \quad \Phi_{12}{}^* = \Phi_{10} \quad \Phi_{20}{}^* = \Phi_{22},$$
$$\Phi_{21}{}^* = -\Phi_{21}, \quad \Phi_{22}{}^* = \Phi_{20}, \quad \Lambda^* = \Lambda. \tag{2.48}$$

Under the Sachs symmetry operation, the equations in our list (2.21)–(2.26) are permuted among themselves and so are those of the lists (2.33)–(2.36) or (2.37), (2.38); so also the commutators [cf. (2.30)-(2.32)]. The Sachs symmetry, together with the prime operation, can be used to simplify the generation of equations; alternatively it provides a useful check of equations obtained by other means.

3. APPLICATION TO 2-SURFACES

In order to establish the connection between our present formalism and existing work which employs an ð operator, we must show that our ð operator does in fact reduce to the one defined previously,[9,10] under the appropriate circumstances. Let S be a spacelike or timelike 2-surface in the space–time M. Then at each point of S there are two preferred null directions, defined by the property that they are orthogonal to S if S is spacelike and tangent to S if S is timelike. We shall discuss only the spacelike case below. The situation for a timelike S is analogous. (Essentially, to pass to the timelike case from the spacelike case we must let l^a and n^a take over the roles of m^a and $\overline{m}^a$; hence, þ and þ′ will take over the roles of ð and ð′ below.)

Let us suppose, then, that S is spacelike and that a null tetrad and dyad system has been set up in M (in the neighborhood of S at least), so that l^a and n^a are perpendicular to S at each point of S. (For this to be possible *globally* over S would require some topological restrictions on S and M—but we shall ignore such matters here.) The freedom in choosing such a tetrad system, lies partly in the fact that it is only the two null *directions* which are uniquely singled out at each point of S, not a complete null tetrad, and partly in the fact that the choice of null tetrad in the remainder of M is (apart from smoothness considerations) quite arbitrary. This latter freedom is of no concern for us here since we shall be interested only in quantities defined at points of S and in differentiations which act within S itself (i.e., in directions tangent to S). The remaining freedom, is, of course, precisely that which the formalism of this paper is designed to handle.

Consider any tensor field defined on M. At each point of S we shall have a uniquely defined tensor within S, which is obtained by projecting the original tensor orthogonally into the surface. Conversely, any tensor defined within the 2-surface S can be obtained by such a projection. In particular, some tensor fields defined on M will have the property that they are completely tangent to S at each point of S. Such fields we regard as being *unaffected*, at points of S, by the above orthogonal projection. We may interpret any tensor field defined *within* S as such a tensor field in M modulo its behavior off S.

To be more explicit about the nature of this orthogonal projection, define

$$E_a^b = -m_a\overline{m}^b - \overline{m}_a m^b \quad \text{and} \quad F_a^b = l_a n^b + n_a l^b. \tag{3.1}$$

Then we have

$$E_a^b E_b^c = E_a^c, \quad F_a^b F_b^c = F_a^c, \quad F_a^b E_b^c = 0 = E_a^b F_b^c,$$
$$\text{and} \quad E_a^b + F_a^b = \delta_a^b. \tag{3.2}$$

Furthermore, if

$$\xi^a = \eta^a + \zeta^a, \tag{3.3}$$

where η^a is the part of ξ^a linearly dependent on m^a and $\overline{m}^a$ and where ζ^a is the part of ξ^a linearly dependent on l^a and n^a, then we have

$$\eta^a = \xi^b E_b^a, \quad \zeta^a = \xi^b F_b^a, \quad \eta_a = \xi_b E_a^b, \quad \zeta_a = \xi_b F_a^b. \tag{3.4}$$

Thus, E_a^b is the orthogonal projector which can be used to project a vector or tensor at a point of S *into* S, while F_a^b can be used to pick out the parts of a tensor which are *perpendicular* to S. For example, if the tensor $T_{ab}{}^c$ on M is defined at a point of S; then the tensor $U_{ab}{}^c = T_{pq}{}^r E_a^p E_b^q E_r^c$, at this point, is the orthogonal projection of $T_{ab}{}^c$ into S. If $U_{ab}{}^c = T_{ab}{}^c$, then $T_{ab}{}^c$ is completely tangent to S at this point. In this case, the projections of $T_{ab}{}^c$ which involve components perpendicular to S, such as $V_{ab}{}^c = T_{pq}{}^r F_a^p E_b^q E_r^c, \ldots, Z_{ab}{}^c = T_{pq}{}^r F_a^p F_b^q F_r^c$, must all vanish. Thus we may interpret $T_{ab}{}^c (= U_{ab}{}^c)$ at the point as a 2-tensor defined within the 2-surface S. For a general $T_{ab}{}^c$ which is not completely tangent to S, there will be some nonvanishing projections among $V_{ab}{}^c, \ldots, Z_{ab}{}^c$. The information contained in $T_{ab}{}^c$ will be shared among all of $U_{ab}{}^c, V_{ab}{}^c, \ldots, Z_{ab}{}^c$; but only $U_{ab}{}^c$ is interpretable as a 2-tensor defined within S.

If we restrict attention for the moment only to tensors, such as $U_{ab}{}^c$, which are completely tangent to S, then since we may regard them as 2-tensors within S (at points of S) we may ask that the definition of *covariant derivative* within S be defined in terms of the four-dimensional covariant derivative operator ∇_a. To do this, we merely apply the four-dimensional operator ∇_a to any tensor which is completely tangent to S and then project the resulting tensor back into S. The resulting tensor, when interpreted as a 2-tensor defined within S is precisely the (two-dimensional) covariant derivative of the original tensor interpreted as a 2-tensor.[13]

For example, in the case of $U_{ab}{}^c$ above, the two-dimensional covariant derivative of $U_{ab}{}^c$ may be interpreted as

$$E_a^p E_b^q E_r^c (E_d^s \nabla_s) U_{pq}{}^r. \tag{3.5}$$

We can also define a "covariant derivative" operation (analogous to Fermi or Fermi–Walker transport,[14] but where the 2-surface S replaces the curve) for quantities defined at points of S which need not be completely tangent to S. For example, in the case of $V_{ab}{}^c, \ldots, Z_{ab}{}^c$ defined above, the respective "covariant derivatives" would be

$$F_a^p E_b^q E_r^c (E_d^s \nabla_s) V_{pq}{}^r, \ldots, F_a^p F_b^q F_r^c (E_d^s \nabla_s) Z_{pq}{}^r. \tag{3.6}$$

Let us now return to consideration of the set of spin- and boost–weighted quantities associated with a tensor. Take the tensor $T_{ab}{}^c$, for example. We can construct a total of 64 spin- and boost-weighted scalars from $T_{ab}{}^c$, namely those quantities obtained from $T_{ab}{}^c$ by transvection with the various vectors of the null tetrad. If we transvect only with the vectors m^a and $\overline{m}^a$ then we obtain eight quantities referring to the part of $T_{ab}{}^c$ which is completely tangent to S. For example, $m^a\overline{m}^b m_c T_{ab}{}^c = m^a\overline{m}^b m_c U_{ab}{}^c$. These quantities may be thought of as arising from a 2-tensor defined within S. Since m^a and $\bar{m}^a$ are of respective types $\{1, -1\}$ and $\{-1, 1\}$, it follows that the type of any such quantity will be of the form $\{s, -s\}$. Thus, the boost weight of such a quantity vanishes and the spin-weight s alone serves to characterize the type.

We have remarked above that the two-dimensional covariant derivative of $U_{ab}{}^c$ may be obtained by projecting the four-dimensional derivative of $U_{ab}{}^c$ back into the surface S. We may achieve this projection instead by contracting with m^a and $\bar{m}^a$. For example,

$$m^a\bar{m}^b m_c(m^d\nabla_d)U_{ab}{}^c = m^p m^q \bar{m}^r m_s E^a_q E^b_r E^s_c (E^d_p\nabla_d)U_{ab}{}^c. \tag{3.7}$$

Expression (3.7) can be rewritten in terms of the $\eth$ operator as follows:

$$m^d m^a \bar{m}^b m_c \nabla_d U_{ab}{}^c = \eth(m^a\bar{m}^b m_c U_{ab}{}^c). \tag{3.8}$$

Perhaps the easiest way to see this is to write the left-hand side as $m^a\bar{m}^b m_c \eth U_{ab}{}^c$—where we allow $\eth$ to act on tensorial quantities (here $U_{ab}{}^c$ has type $\{0,0\}$)—and observe that $E^b_a \eth m^a = 0$ (m^a having type $\{1, -1\}$) and $E^b_a \eth \bar{m}^a = 0$ ($\bar{m}^a$ having type $\{-1, 1\}$). In fact, we have quite generally (allowing $\eth$ and Þ to act on spinor quantities)[15]

$$\text{Þ}o^A = -\kappa\iota^A, \quad \eth o^A = -\sigma\iota^A, \quad \eth' o^A = -\rho\iota^A,$$
$$\text{Þ}' o^A = -\tau\iota^A; \tag{3.9}$$
$$\text{Þ}\iota^A = -\tau' o^A, \quad \eth\iota^A = -\rho' o^A$$
$$\eth'\iota^A = -\sigma' o^A, \quad \text{Þ}'\iota^A = -\kappa' o^A.$$

Thus we obtain particularly,

$$\eth m^a = -\bar{\sigma}' l^a - \sigma n^a, \quad \eth\bar{m}^a = -\rho' l^a - \bar{\rho} n^a. \tag{3.10}$$

Expression (3.8) shows that $\eth$ is, in effect, just a two-dimensional covariant differentiation operator acting within the surface S. In fact (3.8) is a simple extension, apart from the absence of their factor $\sqrt{2}$ on the right-hand side, of an expression given by Goldberg *et al.*[10] [cf. their Eq. (2.13)] for the operator $\eth$ as had been originally defined by Newman and Penrose.[9] We observe that any spin-weighted quantity defined on S (of integral spin weight) can be expressed in the form $m^a \cdots m^c \bar{m}^d \cdots \bar{m}^f T_{a\ldots cd\ldots f}$, so (3.8) fully characterizes $\eth$ on S. To get complete agreement with the Newman–Penrose (θ, ϕ) expressions for $\eth$, we should require S to be intrinsically a metric sphere of radius $2^{-1/2}$;

$$-ds^2 = \tfrac{1}{2}(d\theta^2 + \sin^2\theta d\phi^2) \tag{3.11}$$

on S. But for complete agreement in the case of a *general* 2-surface metric we would have (strictly speaking) to multiply up our $\eth$ operator by a factor $\sqrt{2}$. (It is perhaps unfortunate to have to introduce a discrepancy of this kind into our definition of $\eth$. However, the formulas of this paper would look unnecessarily cumbersome were we to retain complete *numerical* agreement with the original Newman–Penrose definition.)

Note that, as applied to spin-weighted scalars defined on the surface S, the operator $\eth$ is really completely *intrinsic* to S; that is to say, its effect depends only on the intrinsic metric of S and not on the way that S is embedded in the space–time M. When applied to boost-weighted scalars, the effect of $\eth$ *does* depend on the embedding. We can illustrate both these facts if we examine the geometrical meaning of the commutator (2.32). In our present situation the quantities ρ and ρ' are both necessarily real at points of S. This is because l^a and n^a are the null tangents (i.e., normal) to the two null hypersurfaces which intersect in S, so the curl of both l^a and n^a must vanish. (This is just the condition that the plane-elements spanned by the real and imaginary parts of m^a are surface-forming.) Thus, the commutator (2.32) becomes

$$(\eth\eth' - \eth'\eth)\eta = -(pK - q\bar{K})\eta, \tag{3.12}$$

where

$$K = \sigma\sigma' - \rho\rho' - \Psi_2 + \Lambda + \Phi_{11}. \tag{3.13}$$

The complex scalar K (type $\{0,0\}$) is a kind of *complex* curvature for the surface S and has been studied earlier in connection with the characteristic initial value problem.[16] The Gaussian curvature of the surface S is twice[17] the real part of K:

$${}^{(2)}R = (K + \bar{K}).$$

This can be seen by choosing $p = 1$, $q = -1$ in (3.12); then η is defined by a real tangent vector y^a given by

$$y^a = -\bar{\eta} m^a - \eta\bar{m}^a, \quad \text{so } \eta = m^a y_a. \tag{3.14}$$

Equation (3.12) expresses the rotation of y^a as the vector y^a is parallel-transported within S around a small loop on S. In a similar way, if we choose $p = q = 1$ in (3.12), then η is defined by a vector z^a normal to S given by

$$z^a = \xi l^a + \eta n^a, \quad \text{so } \eta = l^a z_a. \tag{3.15}$$

As the vector z^a is transported around a small loop on S [this is the generalized Fermi transport of (3.6)—the vector must be continually projected out perpendicular to S as it is carried around], then it undergoes a boost of magnitude given by the product of $i(\bar{K} - K)$ with the small area enclosed by the loop. Thus the imaginary part of K describes an extrinsic curvature invariant of S in M.

To sum up, we see that our formalism provides an effecive means of studying the intrinsic geometry of Riemannian 2-surfaces. This is when we restrict attention merely to spin-weighted quantities (type $\{s, -s\}$) and employ a tetrad of vectors related to a surface S as described above. This is essentially the same $\eth$ formalism as has been employed previously. However, if we allow spin- and boost-weighted quantities (general type $\{p, q\}$) then we can also use our formalism for the study of the extrinsic geometry of 2-surface embedded in a space–time. (This is still employing only the $\eth$ and $\eth'$ operators at points of a 2-surface S.) But our formalism is, of course, much more general than this since we can use the Þ and $\text{Þ}'$ operators to study the relation between different embedded 2-surfaces, and also in circumstances in which the plane-elements spanned by the real and imaginary parts of m^a need not be surface-forming.

4. CONCLUSIONS

We have presented, here, a formalism which should have many applications to the type of situation in which the NP formalism has proved useful in the past. It is likely that the present approach will provide considerable formal simplifications in most cases. In particular, it should yield asymptotic formulas directly in terms of the familiar ð operator, obviating any necessity of having to pass through a complicated intermediate stage of explicit dependence on angular coordinates. We have contented ourselves here with presenting only the basic formalism. Detailed applications will be given elsewhere.

ACKNOWLEDGMENT

The authors are indebted to Dr. M. Walker and Dr. J. Stewart for pointing out errors in the original manuscript

---

*Supported in part by the National Science Foundation under Grant GP-34721X with the University of Chicago.

[1] Cf. L. Eisenhart, *Riemannian Geometry* (Princeton, U. P., Princeton, N. J., 1966).

[2] R. Penrose, *Battelle Rencontres*, edited by C. DeWitt, J. Wheeler (Benjamin, New York, 1968).

[3] Whenever a Latin letter such as $a, b, \ldots, a_0, \ldots$ or $A, B, \ldots, A_0 \ldots A', \ldots$ is used as an index, it is to be read merely as a *label* indicating the tensor or spinor nature of a quantity. Such indices are treated formally and are not assigned values such as 0,1,2,3. Thus ξ^a is to be thought of *as* a vector, not as a set of components. Similarly, $\psi_A{}^{BC'}$ will denote an actual spinor and not a set of spinor components. When we wish to look at individual components of a vector, or tensor, or spinor we require a basis frame to be introduced. This basis frame could, in the case of tensors, be that defined by some coordinate system (holonomic frame) or it could be introduced independently as a tetrad system, orthonormal or otherwise (nonholonomic frame). For spinors we would require a spinor dyad system. Components in a basis frame are labeled in Ref. 2 by German letters $\mathfrak{a}, \mathfrak{b}, \ldots, \mathfrak{a}_0, \ldots$ (tensor components) or $\mathfrak{A}, \mathfrak{B}, \ldots, \mathfrak{A}_0, \ldots \mathfrak{A}', \ldots$ (spinor components) and for consistency's sake we shall stick to this here. Then $\mathfrak{a}, \mathfrak{b}, \ldots$ range over 0,1,2,3, and $\mathfrak{A}, \mathfrak{B}, \ldots$ range over 0,1 etc. One significant advantage of this type of formulation is that it enables the spinor–tensor correspondence to be expressed without the use of Infeld–van der Waerden connection symbols $\sigma_{\mathfrak{a}}{}^{\mathfrak{A}\mathfrak{A}'}$. It is now legitimate to *equate* a tensor with its spinor equivalent, so we can correctly write $\chi_a{}^{bc} = \chi_{AA'}{}^{BB'CC'}$. The tensor label a may thus be regarded as a *shorthand* for the pair of spinor labels AA'; similarly b is a shorthand for BB', etc. (The symbols $\sigma_{\mathfrak{a}}{}^{\mathfrak{A}\mathfrak{A}'}$ would merely express the relation between tensor and spinor frames.)

[4] E. Newman and R. Penrose, J. Math. Phys. 3, 566 (1962). [In using equations from this paper, we have taken account of the published errata: J. Math. Phys. 4. 998 (1963)].

[5] R. Penrose, "An Analysis of Space–Time," Adam's Prize Essay, 1966.

[6] Although we shall treat the prime as effectively an allowable operation within the formalism, it is possible in certain circumstances to have quantities to which the prime could not be applied (e.g., in an algebraically special vacuum solution, which is not type D, if l^a were chosen in the direction of the repeated principal null direction, then quantities could be definable from l^a having no analog in terms of n^a). When applied to the quantities defined in this paper, however, the prime will denote a well-defined operation.

[7] The symbol Þ is pronounced "thorn" and ð is pronounced "e(d)th"; Þ and ð are the phonetic symbols for the soft and hard "th", respectively.

[8] J. Winicour, private communication, has defined a similar operator ð having some, but not all of the properties of our ð operator.

[9] E. Newman and R. Penrose, J. Math. Phys. 7, 863 (1966).

[10] J. N. Goldberg, A. J. Macfarlane, F. Rohrlich, E. C. G. Sudarshan and E. T. Newman, J. Math. Phys. 8, 2155 (1967).

[11] F. A. E. Pirani, *Brandeis Summer Institute, 1964* (Prentice-Hall, Englewood Cliffs, N. J., 1965).

[12] R. K. Sachs, private communication (1961).

[13] One may verify that this operation is indeed covariant differentiation within S by observing that the various axioms for covariant differentiation are satisfied, with regard to tensors completely tangent to S.

[14] Fermi–Walker transport (A. G. Walker, Proc. R. Soc. Edinb. **52**, 345 (1932)) of a vector v^a with respect to a (timelike) curve γ may be defined in the following way. First we decompose v^a into two parts, $v^a = p^a + t^a$, where p^a is perpendicular to γ and t^a is tangent to γ. The part t^a is to undergo "parallel transport" within γ (i.e., to have merely its length $t^a t_a$ remain constant), while p^a is to undergo Fermi transport [E. Fermi, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Nat. Rend. **31**, 21 (1922)] along γ. The definition of Fermi transport is such that p^a is carried along γ always remaining perpendicular to γ. This is achieved by demanding that the projection orthogonal to γ, of the covariant derivative of p^a in the tangent direction to γ, should vanish. In a similar way, we can define a means of transporting a vector v^a in directions lying within a 2-surface S. Again we form a decomposition $v^a = p^a + t^a$, where p^a is now perpendicular to S and t^a is tangent to S. The vector t^a is to be carried by (intrinsic) parallel transport within S, while p^a undergoes a form of "Fermi transport" on S. Thus, we demand that the projection orthogonal to S, of the covariant derivative of p^a in directions tangent to S, should vanish. However, since S is two-dimensional, this analog of Fermi(–Walker) transport is not normally "integrable." That is to say, the transport of v^a is dependent upon the path within S along which it is carried.

[15] Equations (3.9), (3.10), etc. can be used quite generally to translate covariant equations into our formalism as follows: $m^a \bar{m}^b o^E (\nabla_a T_{bE}) = \bar{m}^b o^E \eth T_{bE} = \eth(T_{bE} \bar{m}^b o^E) - T_{bE} o^E \eth \bar{m}^b - T_{bE} \bar{m}^b \eth o^E = \eth(T_{bE} \bar{m}^b o^E) + \rho' T_{bE} l^b o^E + \bar{\rho} T_{bE} n^b o^E + \sigma T_{bE} \bar{m}^b l^E$.

[16] R. Penrose, "Null Hypersurface Initial Data for Classical Fields of Arbitrary Spin and for General Relativity", in P. G. Bergmann's ARL Tech. Documentary Report 63-56, (1963).

[17] The numerical factor in the equality of $K + \bar{K}$ with the Gaussian curvature of S can be checked by choosing S to be a sphere of radius $2^{-1/2}$ and comparing with Refs. 9 and 10.

[18] For normal applications, p and q will be integers, so $p = \bar{p}$ and $q = \bar{q}$. However, the formalism will still work if p and q are any pair of complex numbers for which $p - q$ is an integer. See, for example, M. A. Naimark, *Linear Representation of the Lorentz Group* (Pergamon, Oxford, 1964).

59

BLACK HOLES

Cosmology Now (Ed. L. John, BBC Publications, London 1973; Taplinger, New York 1976) pp. 103–26. Based on BBC Radio 3 Series; *Albert Einstein: His Influence on Physics, Philosophy and Politics* (Ed. Peter C. Aichelburg and Roman U. Sexl), pp. 33–50, Friedr. Vieweg & Sohn, 1979; *The World Treasury of Physics, Astronomy, and Mathematics* (Ed. Timothy Ferris), pp. 203–25, Little Brown, 1991

610 Black holes. *Cosmology Now* (Ed. L. John, BBC Publications, London 1973; Taplinger, New York 1976) pp. 103–26

The heavy elements of which we are composed were forged in the nuclear furnaces inside giant stars. Yet many of the stars which now contain heavy elements are of the wrong type to have been able to produce them. The elements must have been ejected by a previous generation of supernovae and swept up by the stars of today. But computer studies show that to have produced the abundance of heavy elements we now observe, the mass of 'first generation' stars must have been considerably higher than now. What has happened to the missing stars? Some cosmologists are suggesting that up to 90 per cent of the matter originally present in the universe has disappeared forever down black holes.

In the following chapter, Roger Penrose, Professor of Mathematics, at University of Oxford, shows that these curious cosmological objects have a thoroughly respectable pedigree; one which can be traced in unbroken line to stars not very different from our own. L. H. J.

About 6,000 light years away, in the constellation of Cygnus the Swan, lies the blue supergiant star HDE 226868. Its mass exceeds that of our sun by a factor of about thirty, and its radius by a factor of nearly twenty-five. This in itself is nothing especially unusual. Many other stars of a similar nature are known. But once every five and one-half days, HDE 226868 circles in orbit about an invisible companion. It is this mysterious companion which concerns us here—with a mass one-half that of HDE 226868 but utterly tiny, its radius apparently being only about thirty miles! The companion of HDE 226868 is now believed by many astronomers to be a *black hole*—a bizarre consequence of the physical laws embodied in Einstein's theory of general relativity. Einstein's theory describes gravitation in terms of 'space-time curvature', and in a black hole, the gravitational field has become so strong that this curvature leads to some surprising and weird effects.

The identification of the companion of HDE 226868 as a black hole is not yet quite certain, but looks highly probable at the present time. There are other objects in the heavens which some astronomers claim are likely also to be black holes and it may be that the final definitive judgement on the existence of black holes will be centred instead on one of these. But however it comes, this discovery will be an event of the utmost importance to present-day physical theory. For theory predicts that black holes *should* exist and should occur sometimes as the end-point of stellar evolution. If black holes were found *not* to exist then this would point to some drastic revision necessary in the theory. On the other hand their *existence* will also pose fundamental problems for the theory and I shall attempt to elucidate some of these later in this chapter.

What is a black hole? For astronomical purposes it behaves as a small, highly condensed dark 'body'. But it is not really a material body in the ordinary sense. It possesses no ponderable surface. A black hole is a region of empty space (albeit a strangely distorted one) which acts as a centre of gravitational attraction. At *one* time a material body *was* there. But the body collapsed inwards under its own gravitational pull. The more the body concentrated itself towards the centre the stronger became its gravitational field and the less was the body able to stop itself from yet further collapse. At a certain stage a point of no return was reached, and the body passed within its 'absolute event horizon'. I shall say more of this later, but for our present purposes, it is the absolute event horizon which acts as the boundary surface of the black hole. This surface is not material. It is merely a demarcation line drawn in space separating an interior from an exterior region. The interior region—into which the body has fallen—is defined by the fact that no matter, light, or signal of any kind can escape from it, while the exterior region is where it is still possible for signals or material particles to escape to the outside world. The matter which collapsed to form the black hole has fallen deep inside to attain incredible densities, apparently even to be crushed out of existence by reaching what is known as a 'space-time singularity'—a place where physical laws, as presently understood, must cease to apply.

Since the black hole acts as a centre of attraction it can draw new material towards it—which once inside can never escape. The material thus swallowed contributes to the effective mass of the black hole. And as its mass increases the black hole grows in size, its linear dimensions being proportional to its mass. Its attractive power likewise increases, so the alarming picture presents itself of an ever-increasing celestial vacuum cleaner—a maelstrom in space which sweeps up all in its path. But things are not quite so bad as this. We are saved by the very minuteness of black holes—a fact which results from the smallness of the gravitational constant.

To see this, let us return to our picture of HDE 226868 (Fig. 13). Accepting the most recent figures for the dimensions involved, we have a black hole of some thirty miles in radius—in mutual orbit about a giant star whose radius is over 300,000 times larger. Despite its small size, the gravitational influence of the black hole is sufficient to distort the large star considerably out of spherical shape. It becomes rather like an egg whose small end is somewhat pointed in the direction of the black hole. A certain amount of material is dragged from this point and slowly falls inward to the black hole. It does not fall straight in, however. The black hole behaves much like a point mass. Most of the material dragged from the large star will remain circulating about the black hole for a long time. Only gradually, as frictional effects begin to play their part, will the material begin to spiral inwards. Again we must bear in mind the small size of the hole. (Imagine having to drain a normal-sized bath through a plughole a ten-thousandth of an inch across—or a bath the size of Loch Lomond through a normal-sized plughole!) The material can be only very slowly funnelled into the black hole. And as it gets funnelled in it gets compressed and very hot—so hot that the material must be expected to radiate light of very short wavelength, X-rays, in fact. Such X-rays are actually observed coming from the vicinity of HDE 226868. And the source of these X-rays (referred to as Cygnus X-1) appears, on the basis of detailed observations, to be in orbit about the visible component HDE 226868. The observed signals seem to be perfectly consistent with the black-hole picture I have presented. However we should remain cautious about drawing premature conclusions, as it is still conceivable that some alternative explanation of the observations may eventually turn out to be correct. The present evidence seems to be pointing ever more strongly in favour of Cygnus X-1 being a black hole, but even if for some reason this interpretation does turn out to be erroneous after all, it would still be very surprising

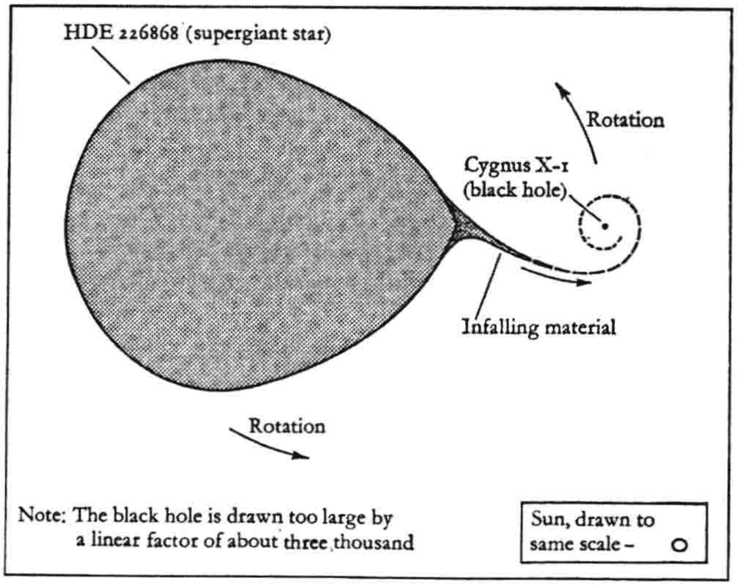

HDE 226868 (supergiant star)
Rotation
Cygnus X-1
(black hole)
Infalling material
Rotation
Note: The black hole is drawn too large by
a linear factor of about three thousand
Sun, drawn to
same scale –

Fig. 13: The gravitational field of the black hole distorts the supergiant star out of spherical shape and drags material from it.

(on the basis of present theory) if *no* black holes were found to exist. To indicate why, I should explain something of the picture that astronomers and astrophysicists have developed concerning stellar evolution and then indicate some of the theory that lies in support of the black hole picture that I have presented.

Let us consider first what theory and observations tell us to expect for the future of our sun—or of any other normal star of about the same mass. After shining at approximately its present brightness for about 7,000 million years the sun will begin a change which will transform it beyond recognition. According to the well-accepted theory of stellar evolution the sun will grow to an enormous size and become, like stars such as Antares in the constellation Scorpio, a red giant some 200 million miles in diameter. By this time the planets Mercury, Venus and the earth will have been burned away and their former orbits will lie well within the new solar surface. The density of the sun's material will by then have fallen from its present value of a fifth that of the earth to a tenth of the density of air!

As it continues to burn more and more of its nuclear fuel, making heavy inroads on its helium and heavier elements as well as hydrogen, the bloated sun will halt its expansion and begin to contract—down past its present size, smaller and smaller until it stabilises as a white dwarf star perhaps about the size of the earth. At this stage further contraction will be impossible because the electrons of its atoms will be packed together so closely that a law of quantum mechanics known as Pauli's Exclusion Principle will come into play. This principle states that no two electrons in an atom can occupy the same energy state. One can envisage the atoms so closely squeezed together under the dwarf sun's tremendous gravitational field that, pictorially speaking, any closer spacing would force the electrons to get in each other's way so that they cannot, at these temperatures and pressures, be squashed together any further. In this state the density of the solar material will be such that a matchbox full of it would weigh several tons. No material on earth has a density remotely approaching that of a white dwarf but as with red giants, many white dwarfs can be seen in our galaxy. Their ultimate fate is simply to cool off to form black dwarfs and thereafter act merely as centres of strong gravitational attraction. The planets, Jupiter, Saturn, Uranus, Neptune, Pluto and possibly Mars, will still continue to circle the ancient sun aeons after it has died.

White dwarfs are part of the normal evolutionary history of average-sized stars like the sun, and astronomical observations show actual stars at each stage of stellar evolution through the stage now reached by our sun, on to the red giant phase and back to white dwarfs. Moreover the theory of stellar physics fits these observations closely, but not all stars can follow this 'normal' evolutionary path.

As long ago as 1931, Subrahmanyan Chandrasekhar calculated that there must be a maximum mass above which a white dwarf could not sustain itself against even further gravitational contraction despite Pauli's Exclusion Principle. Furthermore, this mass was not much greater than the mass of the sun—1.4 times according to Chandrasekhar; a little less according to more recent computations. Many stars have masses considerably more than 1.5 times that of the sun. What is going to happen to them?

The answer depends on just how heavy the star is. Consider a star of twice the mass of the sun. Like the sun it will also expand to an enormous size and then recontract, but being more massive than Chandrasekhar's limit for a white dwarf it will be unable to settle down to final equilibrium in the white dwarf state. To picture what happens it will be useful to consider the giant phase of a star more fully. As soon as the central density of the star reaches that of a white dwarf, the outer layers of the star expand, and they go on expanding as more and more of the central material gets compressed into a white dwarf state. So the giant star develops a growing white dwarf core. In the case of the sun, all the material that remains in the star will eventually become part of this white dwarf. But if the star is too massive, there comes a point at which the core effectively exceeds Chandrasekhar's limit, whereupon it promptly collapses. In the process of collapse there is a tremendous release of energy, much of which is in the form of neutrinos which are absorbed (so it is believed) in the outer regions of the star, heating the envelope to an enormous temperature. A cataclysmic explosion ensues—a supernova explosion—which blows off a considerable proportion of the mass of the star. Many supernovae have been seen, the last one in our galaxy being described by Johannes Kepler in 1604. A supernova will outshine a whole galaxy for a few days and may even be visible by daylight. As much as 90 per cent of the star's mass might be thrown off by the cataclysmic explosion to form tenuous debris of heavy elements which may later enrich the stars of a second generation. But especially interesting is the collapsed remnant of the star left behind at the centre of the rapidly expanding cloud of ejected gases. This core is much too compressed to form a white dwarf and it can only find equilibrium as a neutron star.

A neutron star is tiny even by comparison with a white dwarf. Indeed the reduction from a white dwarf to a neutron star is even more than the reduction of 100 to one from the sun to a white dwarf, or probably rather more than the approximate reduction of 250 to one from a red giant to the sun. A neutron star may be only 10 kilometres in radius or only about one seven-hundredth the radius of a white dwarf. The density of a neutron star could be more than a hundred million times the already extraordinary density of a white dwarf.

Our matchbox full of neutron star material would now weigh as much as an asteroid a mile or so in diameter. The star's density would be comparable with the density of the proton or neutron itself—in fact a neutron star could in some ways be regarded as an over-sized atomic nucleus, the only essential difference being that it is bound together by gravitation rather than by nuclear forces. Individual atoms have ceased to exist as such. The nuclei are touching and form one continuous mass. As for the electrons, as Professor Lynden-Bell has explained in an earlier chapter, what has happened is that the stupendous gravitational forces have squeezed the electrons into the only space available to them —that already occupied by the protons, reversing the usual reaction so that the star is now composed mainly of neutrons and it is the Pauli Exclusion Principle acting on these neutrons that supplies the effective forces preventing further collapse.

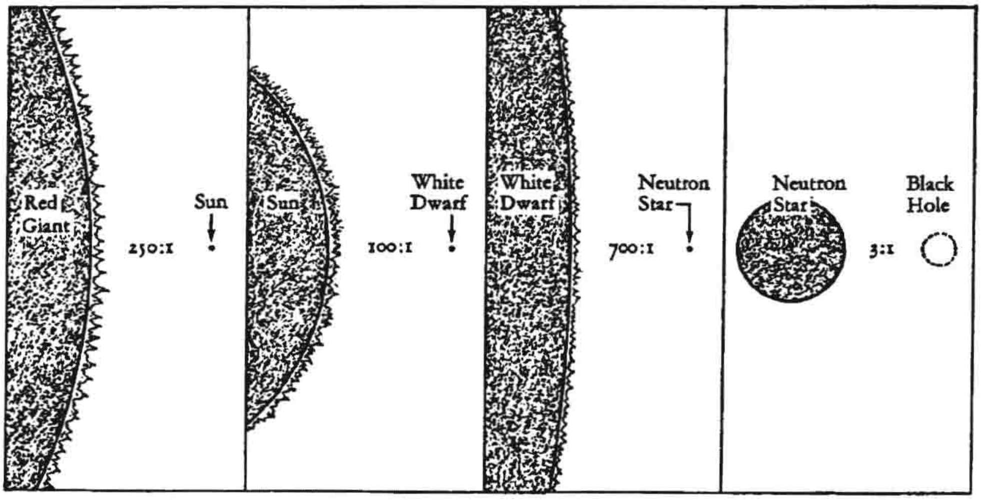
This picture of a neutron star was predicted theoretically by the Soviet physicist Lev Landau in 1932 and studied in detail by J. Robert Oppenheimer, Robert Serber and George M. Volkoff in 1938 and 1939. For years many astronomers doubted whether neutron stars could actually exist. Since 1967 the observational situation has changed dramatically, because in that year the first pulsars were observed. Since then the theory of pulsars has developed rapidly and it now seems virtually certain that the radio and optical impulses emitted by pulsars owe their energy and extraordinary regularity to the presence of a rotating neutron star. At least two pulsars reside inside supernova remnants, one being the Crab Nebula, and this gives further support to the theory that pulsars are in fact neutron stars.

There is a maximum mass above which a neutron star would not be able to sustain itself against still further gravitational contraction. There is some uncertainty as to the exact value of this maximum-mass limit. The original value given by Oppenheimer and Volkoff in 1939 was about 0.7 times the solar mass. More recently, larger values of up to three solar masses have been suggested. These higher values take into account the idea that the heavy subatomic particles called hyperons may be present in addition to ordinary neutrons and protons. But there are stars whose mass is more than fifty times the mass of the sun. What will happen to these? It seems exceedingly unlikely that all these stars would, as a result of their final collapse phase, or earlier, inevitably throw off so much of their material that their masses would always fall below the limits required for a stable white dwarf or neutron star to be the result. In these the neutron core would be unable to remain in equilibrium and would have to collapse further inwards. But what other forms of condensed matter might be possible considering that here we have densities in excess even of the fantastic value that is maintained inside a neutron star?

In this case, theory tells us a different story: although greater densities can be achieved, it is not possible to obtain any further stable final equilibrium configurations. Instead the gravitational effects become so overwhelming as to dominate everything else. Newtonian gravitation theory becomes quite inadequate to handle the problem, and instead, we must turn to Einstein's theory of general relativity. But in so doing, we are led to a picture so strange that even the phenomenon of the neutron star must seem commonplace by comparison. This new picture is the one which has now earned the description of a black hole.

Briefly, a black hole is a region of space into which a star (or collection of stars or any other bodies) has fallen, but from which no light, matter or signal of any kind can escape. Before examining this picture in some detail, consider (Fig. 14) the degree of further contraction that would be necessary for a neutron star to be compressed down to the

Fig. 14: The neutron star illustrated has the same mass as the sun but has a diameter 700 times smaller than the earth.



size of a black hole. We have already seen that from the sun's dimension down to that of a neutron star involves a contraction of about 70,000:1 in linear size; from a red giant to a neutron star, a linear contraction of about 20,000,000:1. In view of this, it is perhaps surprising that a further contraction of only about 3:1 is required for a neutron star of one solar mass to become a black hole—in this case, with a radius of about two miles. Larger black holes are also possible, the radius of the hole being proportional to the mass, so they would result from the collapse of a body more massive than the sun. For example, Cygnus X-1 appears to be about 15 times more massive than the sun, so its radius (assuming little rotation) is about 30 miles.

The reason I emphasise the slight nature of this further contraction, only 3:1, is that faced with the unsettling nature of the black holes, people have naturally asked whether our physical theories are tenable under these extreme conditions. But these theories seem to have worked well in describing a very large range of stars of enormously different sizes and densities. In any case, the conditions under which a black hole is formed are not so extreme as all that—not necessarily more extreme than the situation of a neutron star. For example, the densities involved as the collapsing star crosses the absolute event horizon are not vastly different from those inside a neutron star. The larger the collapsing mass, the less would be this density—less in inverse proportion to the square of the mass. In the case of Cygnus X-1 this density would be rather less than that of a neutron star. It has often been considered by astronomers that collections of mass of up to 100 million suns or more might be involved in gravitational collapse in galactic centres. The density at the time such a huge mass crosses the event horizon might then be only about that of water. So the local conditions need not be excessive when a black hole is formed and there seems no reason to suppose that the black hole condition might render general relativity somehow inapplicable.

On the other hand it must be admitted that general relativity plays anything but an irreplaceable role in observational astronomy. It is still possible that general relativity might be wrong; after all, the experimental tests of general relativity which have been successfully performed are still not very numerous. Although no conflict between the theory and observations exists at present (any apparent conflict being explicable by other means), the observations do not point *conclusively* in the direction of general relativity. There is still scope for alternative theories of gravitation.

But no one would deny that general relativity *is* an excellent theory; almost certainly the most satisfactory theory of gravity available to us. Furthermore the theory which at the present time may be regarded as general relativity's most serious rival (namely the Brans-Dicke-Jordan scalar-tensor theory), itself leads to a black hole picture nearly identical to that arising in Einstein's theory. Even in Newtonian theory a phenomenon similar to that of a black hole may be said to occur. As long ago as 1798, Pierre Simone Laplace used Newtonian theory to predict that a sufficiently massive and concentrated body should be *invisible* because the escape velocity at the surface could be greater than the speed of light. A photon or particle of light emitted from the surface of the body would simply fall back to the surface, it would not escape to be observed at large distances from the body. This description is perhaps arguable, but it shows that there is a situation to be faced, even in Newtonian theory.

I only make these observations to show that if one's sole motive for casting doubts on the tenability of general relativity is to get out of the black hole situation, one might just as well stick with relativity, because its rejection will not necessarily make the black hole go away. So I propose, from here on, to restrict my discussion to considerations solely within the bounds of Einstein's general theory of relativity.

To begin with, consider the standard picture of a non-rotating black hole. The black hole is characterised by a spherical surface whose radius is proportional to the hole's mass. This surface is called the 'absolute event horizon'. Its defining property is that signals emitted inside it cannot escape, whereas from any point outside it signals can be emitted that either do escape or fall into the black hole. The radius of the event horizon can be calculated by multiplying twice the mass by the universal gravitational constant and dividing the result by the speed of light squared. Performing this calculation for the sun yields the result that the sun would have to be collapsed into a sphere four miles in diameter if it were to form a black hole. The absolute event horizon would be the surface of this four-mile sphere.

The body whose collapse is responsible for a black hole's existence has fallen deep inside the event horizon. The gravitational field inside the event horizon has become so powerful that even light itself is inevitably dragged inward regardless of the direction in which it is emitted. Outside the event horizon light escapes if it is aimed suitably outward. The closer the emission point is to the event horizon, the more the wave front of the emitted signal is displaced back toward the centre of the black hole. We may intuitively regard this displacement as being caused by the effect of the gravitation on the motion of the light. Owing to the intense gravitational attraction of the black hole, light travels more easily in the direction of the black hole than in the outward direction. Inside the event horizon the inward pull is so strong that outward motion of light has become impossible. This applies not just to light but to any signal originating within the black hole. As for a photon emitted radially outward from the surface of the black hole, on the event horizon, it will mark time, forever hovering in the surface itself at the same distance from the centre of the black hole.

It may seem that this is odd physics indeed, quite unlike the normal situation in relativity theory, where the speed of light has always the same constant value in all directions. But, strange as it may seem, the local physics in the neighbourhood of the absolute event horizon *is* the same as elsewhere. An observer at the event horizon who tries to measure the speed of light must himself be crossing the horizon by falling inwards into the hole. To him the speed of the light hovering on the horizon is indeed the same constant value, in the outward direction.

It would be natural for a reader who is not familiar with general relativity theory to find such a situation confusing. This is partly because so far we have been using a purely spatial description rather than a space-time one—and for many purposes a space-time picture is more illuminating than a spatial one. Strictly speaking, a space-time picture needs to be drawn in four dimensions, but an overall description of the space-time situation can be obtained by suppressing one of the spatial co-ordinates in the space-time diagram and substituting a time co-ordinate. This gives an instantaneous picture of what is going on at all times, and obviates the need for many sequential 'snapshots' of a developing situation.

Consider a light flash emitted in all directions from a given point in ordinary space. The wave front of the flash would be a sphere centred on the emitting point and growing larger each moment at the speed of light. A purely spatial representation of the flash would be a sequence of spheres (Fig. 15) each sphere larger than the preceding one, marking the position of the light flash's spherical wave front at a given moment in time. A space-time representation of the light flash, however, would be a cone whose vertex represents the time and place at which the light flash is emitted, the cone itself describing the history of the light flash.

By the same token the history of a star's collapse down to a black hole can best be depicted in a space-time representation (Fig. 16). The locations of the light cones at various points in space-time show how light signals propagate in the gravitational field. At some points the light cones are drawn as being tipped over, but this is not something that would be noticed by a local observer. Such an observer would follow a path in space-time that proceeds into the interior of the light cone; his speed can never be greater than the locally measured speed of light, and only inside the light cone is this criterion met. But the tipping of the light cones does affect what an observer at large distances can see. Fig. 16 shows that material particles and light signals which originate inside the event horizon are inevitably driven further inwards. For a particle or signal to cross the event horizon from the inside to the outside it would have to violate the condition mentioned above; it would have to exceed the local light speed, in violation of relativity.

By taking a horizontal section through the space-time picture we get a spatial representation of the situation as in

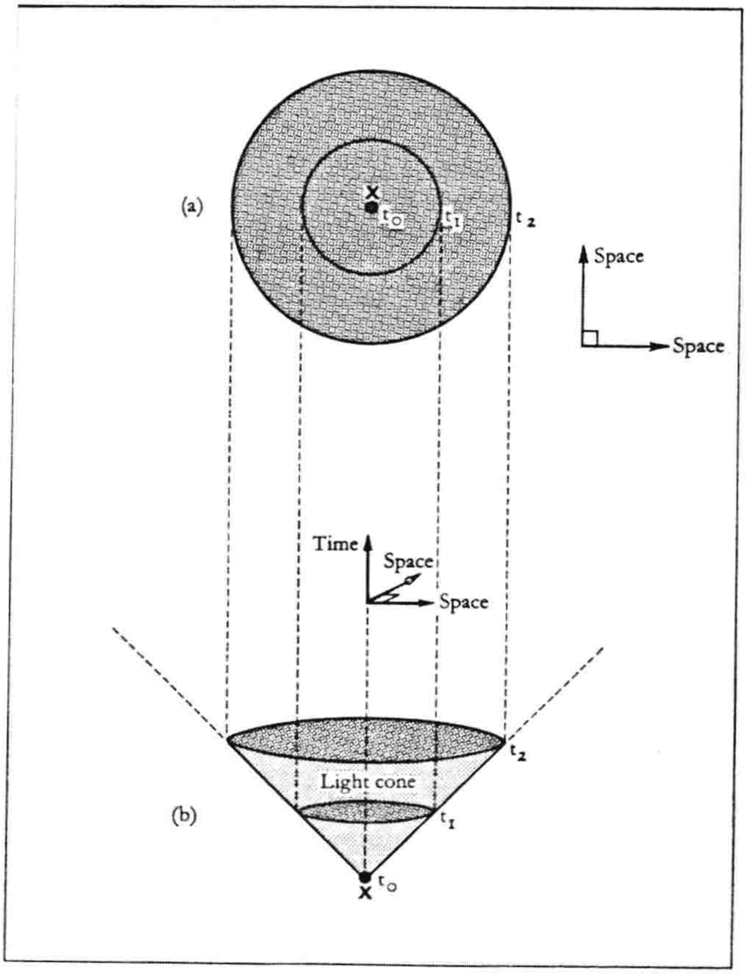


Fig. 15:
(a) propagation of light from x *in space at times* t_1 *and* t_2*.*
(b) propagation of light in space-time to give a 'light cone'.

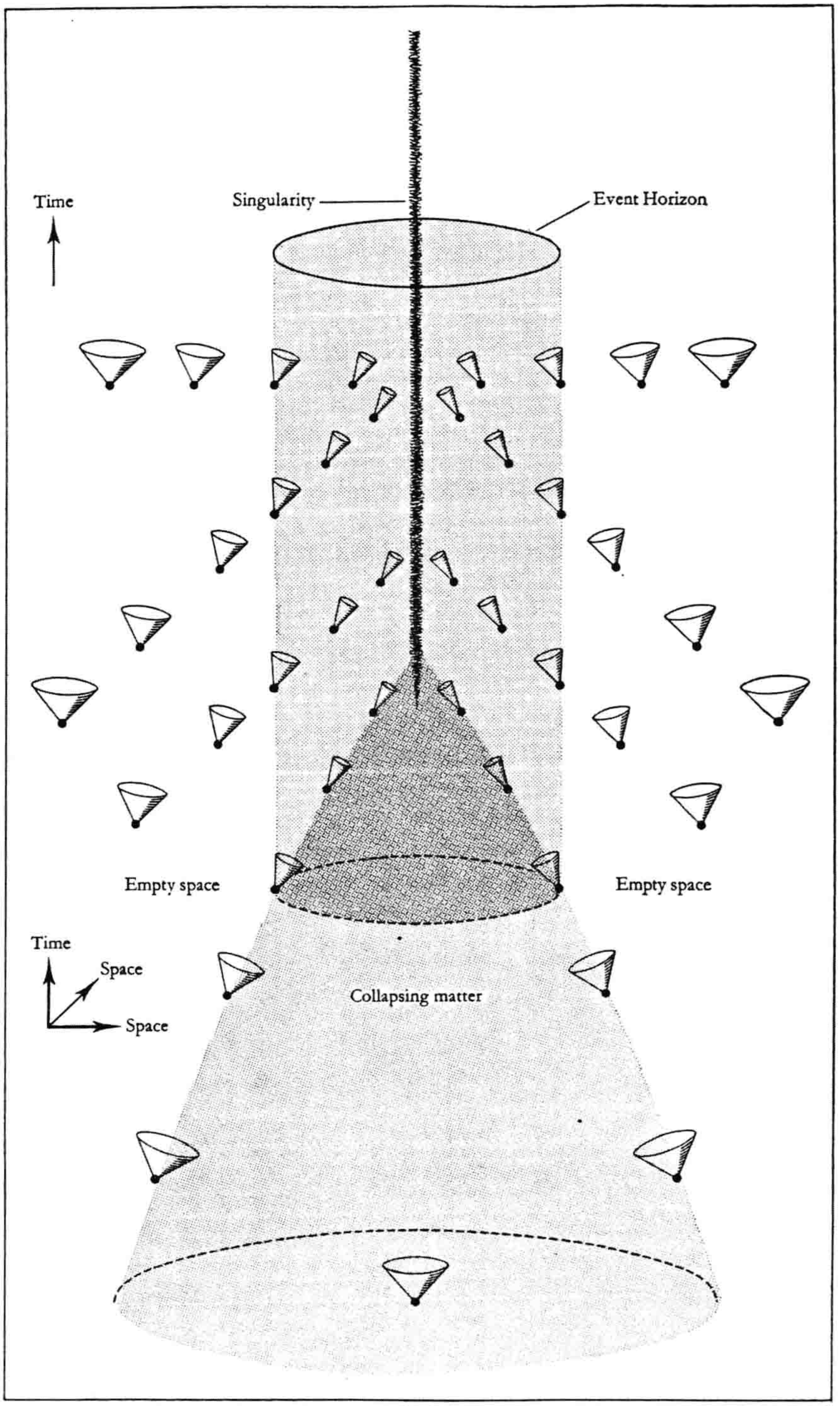

Time
Singularity
Event Horizon
Empty space
Empty space
Time
Space
Space
Collapsing matter

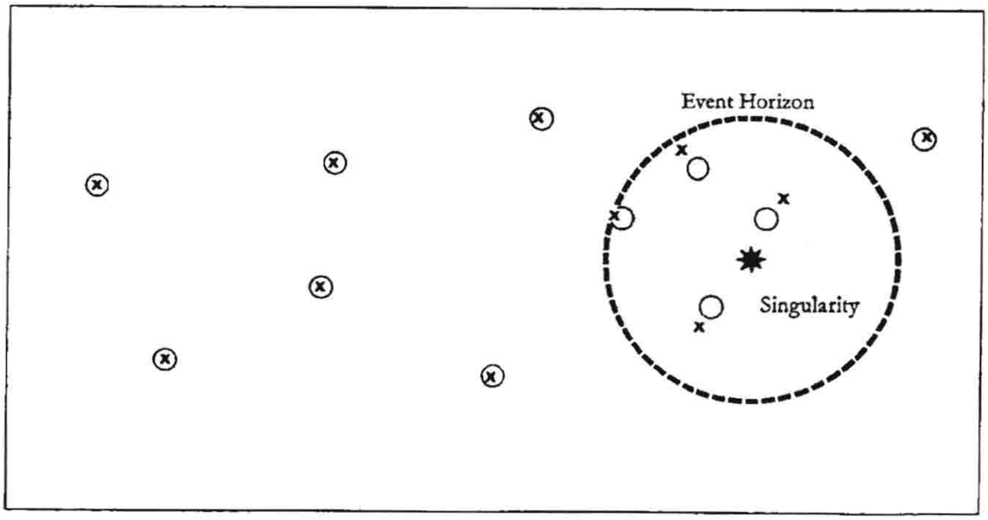
Fig. 17. Perhaps the main advantage of a spatial representation, apart from its greater familiarity, is that one need not dispense with the third spatial dimension when it is important not to do so. The light cones can be depicted as points (the origin of a light flash) surrounded by sphere-like surfaces (the location of the light flash a moment later)—except that when the light cone is tipped over, these surfaces do not actually surround the point of origin. In this later case it would be necessary to exceed the local light speed in order to 'stay in the same place'. A serious drawback with such spatial pictures, therefore, is that it becomes hard to interpret situations of this kind. If a space-time description is used, then it becomes easier to accept that the local physics is the same whether or not the light cones are depicted as tipped over, the 'tipping' being merely an aspect of the local description. (Imagine the figure drawn on a rubber sheet; then the light cone could be twisted around vertically without disturbing much of the rest of the picture.)

Although nothing can ever get out of a black hole, things can fall in. Indeed, it is quite possible that stellar astronauts traversing the depths of space in ages to come will run precisely this risk. Not that they will be likely to encounter a black hole by accident—the smallness of black holes compared with the vastness of the universe will see to that. Indeed, they would have to seek out a black hole deliberately if they wished to experience this 'ultimate trip', and this might be difficult because black holes are not directly visible.

And what will happen to a hapless astronaut who falls into a black hole? What, indeed is the fate of the original body which collapsed to produce the black hole? Assuming that the exact spherical symmetry is maintained right down

Fig. 16 (left): Space-time representation of collapse of a spherically symmetrical body to form a black hole, showing propagation of light inside and outside the event horizon.

Fig. 17: Light propagation inside and outside a non-rotating black hole.



to the centre, the answer provided by general relativity is an alarming one. According to the theory, the curvature of space-time increases without limit as the centre is approached. Not only is the material of the original body squeezed to infinite density at the centre of the black hole—and crushed, effectively, out of existence—but also the vacuum in space-time which is left behind by the body, itself becomes infinitely curved. The effect of this infinite curvature on an observer, were he foolhardy enough to follow the body inwards, would be that he experiences mounting gravitational tidal forces—tidal forces that mount rapidly to infinity.

The gravitational tidal effect is the most direct physical manifestation of space-time curvature. Einstein pointed out that the gravitational force on a body can be eliminated at any one point simply by choosing a frame of reference that is falling freely. He gave the famous example of a lift that broke its cable and fell toward the earth. Any passenger inside would be falling at the same rate as the lift, so he would feel no net gravitational force relative to the lift, and, indeed, would float free of gravity inside it. Such elimination of the gravitational force by free fall is now a familiar feature of space travel. The tidal effect, however, cannot be so eliminated and is, therefore, an absolute manifestation of the gravitational field. Imagine an observer falling freely in the earth's field. Suppose he is surrounded by a sphere of particles that he initially observed to be at rest with respect to himself. The Newtonian gravitational field of the earth varies as the inverse square of the distance between it and any other body, pulling more strongly on objects closer to its surface than on objects farther away. This non-uniformity of the gravitational field tidally distorts the sphere of particles into an ellipsoid which bulges out in the direction of the earth and in the direction away from it. The earth's ocean tides are a familiar example of an effect of this kind, where in this case the earth is experiencing the tidal effects of the moon and the sun.

Fortunately for us, the tidal effects due to gravity encountered in the solar system are small. Nobody complains that his feet experience more gravitational pull towards the earth than his head. There is a difference, but it is not noticeable in the ordinary way. The space-time curvature responsible for this difference has a radius of about the distance from the earth to the sun. (This is a pure coincidence since the sun is itself irrelevant to this particular tidal effect.) At the surface of a white dwarf, on the other hand, the space-time curvature is considerably larger, the radius of curvature being of the same order as the radius of the sun.

This tidal effect would be very noticeable to an astronaut in orbit around the white dwarf. In fact, his head and feet could experience a difference in forces of perhaps one-fifth of the total force the astronaut normally experiences standing on the earth's surface. At the surface of a neutron star, however, the tidal effect is, by ordinary standards, enormous. The radius of space-time curvature there is only about thirty miles. It is clear that no astronaut in a low orbit around a neutron star could possibly survive. For even if he curled himself into a small ball, the gravitational acceleration at various parts of his body would differ by several million times the gravity at the earth's surface.

However, instruments could in principle be built to withstand such high tidal forces—all that would be necessary would be to make them so tiny that the difference between the gravitational attraction on the side nearest the neutron star and the side furthest away is small.

Suppose our astronaut carries such a tiny, rugged instrument as he flies towards a black hole of one solar mass. Long before he reaches the event horizon he will be destroyed by the tidal forces, but his instrument will survive intact as it crosses the event horizon, where it experiences tidal forces about thirty times those at the surface of the neutron star. But it survives only for a while, because the tidal forces encountered in a black hole rise to infinity at the centre. As the instrument falls in towards the centre, the mounting tidal forces will rise rapidly, ripping to pieces in turn the material of the instrument, the molecules of which this material is composed, the atoms which constitute these molecules, the atomic nuclei, and finally the fundamental particles which a moment ago had been the building-blocks of these nuclei. And the entire process would not last more than a few thousandths of a second!

So anything falling into a black hole, whether it be a space ship, a hydrogen molecule, an electron, radio-waves or a beam of light can never emerge again. So far as our universe is concerned it disappears completely and forever into nothing. But how can this be? Is it not a basic law of nature that matter or energy can never be completely destroyed but only converted from one form into another? The question is a perfectly respectable one, but it can be shown by rigorous argument, based on general relativity, that there must be a region inside a black hole, a region of infinite curvature, called a space-time singularity at which the known laws of physics break down. So there is no known conservation law that can be relied on at the centre of the black hole. There is no reason known why the matter should *not* just be totally destroyed as it reaches the singularity. Eventually, perhaps, laws of nature may be formulated which govern the behaviour of space-time singularities, but no such laws are known at present.

This situation has a more familiar manifestation in a somewhat different context. General relativity, like virtually all viable physical theories, is reversible in time. So corresponding to any solution of the equations in which time runs one way, there must be another in which the time-sense is reversed. This leads us to expect that the above situation could—in principle—exist in a time-reversed form. Initially there would be the space-time singularity. Then matter would appear: elementary particles, light. Only later would these particles collect together into atoms, molecules or stars. In fact, a picture of this kind has been considered for many years by astronomers and cosmologists as a model of the creation of the universe. The initial big bang of the cosmological models is, like the centre of a black hole, also a space-time singularity, where the curvature of space-time becomes infinite. But now, rather than being destroyed, matter is created at the singularity. The cosmological big bang is not precisely the time-reverse of a black hole, however, since the singularity is all-embracing, unlike the relatively localised singularity inside the black hole. The basic difference is one of size, and we may indeed envisage more localised 'little bangs', called *white holes*, which are more precisely the time-reverses of black holes. A number of theoreticians have considered such white holes seriously in connection, in particular, with models for quasars. However, I must say that I personally regard the possibility of the existence of white holes with considerable unease. The reason is basically this. Once a black hole is formed there is apparently no means of destroying it. It is created violently, but then settles down and sits around forever—or until the universe re-collapses at the end of time. Now a white hole—the time reverse of a black hole—would have had to have *been* there since the beginning of time—tamely and invisibly biding its time before making its presence known to us. Then, when its moment arrives, it explodes into ordinary matter. But this moment is of its 'own' choosing, governed, apparently, by no definite law. Of course, there is no fundamental reason why this should not occur; the idea simply seems untidy, at variance with thermodynamics and probably also with observation. Nevertheless we *are* still stuck with the big bang and that also seems untidy. But here there appears to be no way out.

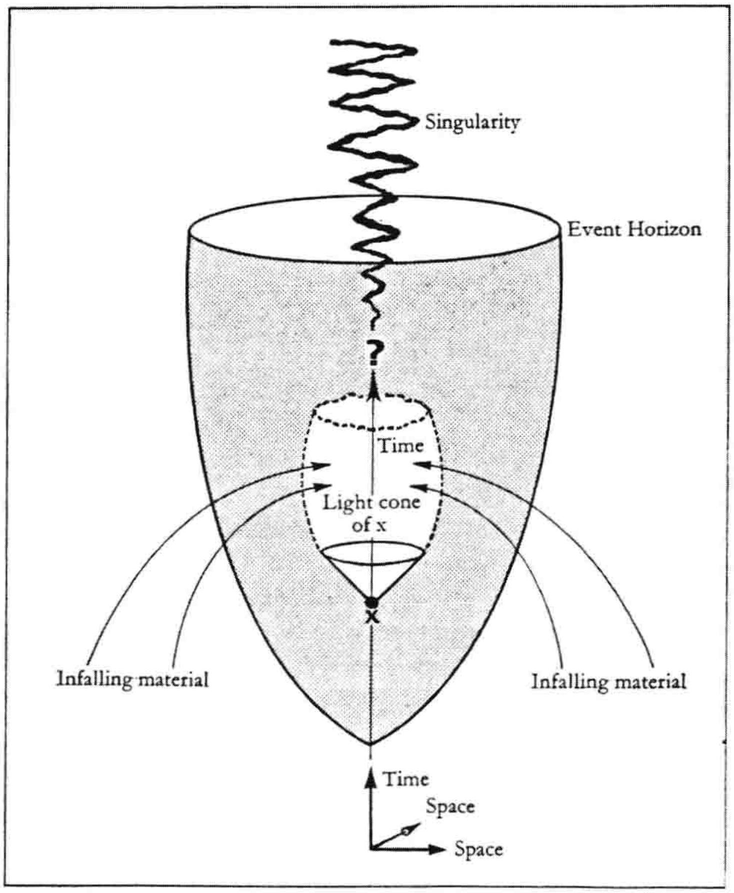
But let us now return from the abyss of speculation and pursue the argument concerning black holes. Quite apart from the doubts I have already raised about the validity of the general theory of relativity, there are other questions that need to be settled before one can fully accept the theoretical concept of the black hole as a realistic description of something that actually occurs in nature. In the first place, can we be sure that enough is known about the nature of matter under the extreme conditions required to form a black hole for the predictions to carry conviction? What role does the assumption of exact spherical symmetry play in the discussion? To what extent does the black-hole picture fit in with astronomical observations? Let us consider these questions in turn.

As I have already pointed out, the densities involved in the formation of a black hole need not be excessive. The same applies to space-time curvatures. A black hole with a mass ranging from 10,000 to 100 million solar masses is sometimes considered a possible candidate for what might inhabit the centre of a galaxy, and a collapsing mass equal to 100 million solar masses would reach a black hole situation when its average density was roughly the density of water. The tidal effects at the event horizon are, like the density, proportional to the inverse square of the mass of the hole, so for a body of 100 million solar masses these tidal effects would be somewhat less than those produced at the earth's surface. As an astronaut passed through the absolute event horizon he would notice nothing. He would have no means of telling that an irretrievable situation had developed, because the exact location of the horizon is not something that can be discerned by local measurement. After this he would have but a few minutes to enjoy the experience of life inside a black hole before the tidal effects mounted to infinity. In the case of a black hole of 10 thousand million solar masses he would have about a day.

The question concerning the role played by the assumption of spherical symmetry has to be examined more carefully. If we do not assume spherical symmetry, then we cannot appeal to the exact solutions of Einstein's equations on which we have based the foregoing discussion. Furthermore, even if we assume that initially the deviations from spherical symmetry are slight, we should have every reason to expect that near the central point these asymmetries would be enormously magnified. Might not the different portions of the collapsing body miss one another? Perhaps they could re-emerge after a close encounter and bounce out again. Even if they did not, can we say anything about the final configurations of the gravitational field resulting from the collapse? It is fortunate that, owing to some general theorems that have been proved over the past few years, a remarkably complete picture of asymmetrical collapse has emerged.

Considering the picture in a little detail, suppose that a massive star or a collection of bodies collapses and that deviations from the spherical symmetry are at first comparatively small. We can establish that a point of no return has been passed if a certain criterion is satisfied. That criterion can be stated in several different ways, but the following is the simplest. Imagine that a flash of light is emitted at some instant at some point in space. The flash of light will follow the light cone centred on the point according to our space-time representation (Fig. 18). The light rays start out from the point by diverging in all directions. When they pass through matter or through a gravitational field, the matter or the field has a focusing effect on the rays. If enough matter or a sufficiently strong gravitational field is encountered, the amount that the rays diverge can be reduced to such an extent that this divergence is actually reversed, that is, the rays start to converge. The required

Fig. 18: The birth of a black hole.



criterion for a point of no return is that every light ray from the space time point encounter enough matter or gravitation for the light cone to be reconverged. It is not hard to show from simple order of magnitude estimates that, for sufficiently large collections of mass, the criterion can indeed be satisfied before densities or curvatures became excessive, and without any assumption of symmetry.

Once this criterion has been satisfied, then according to a precise theorem in general relativity put forward by Stephen Hawking and myself, it follows that there must be a space time singularity somewhere. The theorem does not say that this singularity is necessarily of the same character as that encountered in the centre of a spherically symmetrical black hole, but it is hard to avoid the inference that tidal effects which approach infinity will occur, producing a region of space time where infinitely strong gravitational forces literally squeeze matter and photons out of existence.

Physicists are unhappy with a theory that predicts the evolution of such a truly physical singular state. In the past whenever a singularity was encountered in a theory, it was generally a warning that the theory in its present form was breaking down and new theoretical tools were needed. In the case of black holes we theoreticians are again being presented with a situation of this kind, but one more serious than before, because here the singularity refers to the very structure of space and time.

There are two distinct possibilities at this stage. It may be that the resulting singularity is such that signals can escape from it which can be observed at large distances. This is the more alarming of the two possibilities and it is also the more conjectural. Such a singularity is called 'naked'. The possibility of naked singularities is alarming because the physical effects of near-infinite space time curvatures are quite unknown. If these effects can influence the outside world, then an essential uncertainty is introduced into present physical theory.

On the other hand it is possible that the singularities resulting from gravitational collapse are always hidden from view, as was the case in the spherically symmetrical situation considered above. In that case this essential uncertainty will not arise. This is the hypothesis of 'Cosmic Censorship', according to which naked singularities are forbidden, each singularity being necessarily clothed by an absolute event horizon. There is perhaps some slight theoretical evidence in favour of Cosmic Censorship, but it is only very slight. After all, the consequences of the big bang singularity were decidedly *not* hidden; here was the biggest naked singularity of all time. However, it would be inaccurate to think of the big bang as a violation of Cosmic Censorship. We are concerned here only with singularities which arise in the collapse of perfectly reasonable non-singular matter. I would personally tend to believe that in situations which do not differ too much initially from that of spherical symmetry, the Cosmic Censorship principle is valid. But in more extreme cases the question is, to my mind, quite open. Perhaps there is even observational evidence against Cosmic Censorship. This is a matter to which I shall return later.

If we assume the Cosmic Censorship hypothesis is true, then once the focusing criterion is satisfied an absolute event horizon must arise. This horizon will have a well-defined cross sectional area which will have a tendency to increase with time (black holes can grow but never shrink) but it seems reasonable to suppose that a black hole, left to itself, will settle down to a stationary state. However, we must be cautious in our use of intuition. It might also seem reasonable to suppose that given the vast range of structures, configurations and complexities of the bodies which could have collapsed to a black hole in the first place, the configuration of the black hole itself could also be complex. Some remarkable work by Werner Israel, Brandon Carter and Stephen Hawking has shown that this is probably not the case. Only a very restricted class of stationary black hole configurations can arise. They are uniquely characterised by the value of the mass, spin and charge of the hole. Einstein's equations for the general theory of relativity have been solved explicitly for this problem by Roy P. Kerr and the solution generalised to include charge by Ezra Newman and his co-workers. The reason the asymmetries present in the collapsing body do not show up in the final state of the black hole is that once the hole is formed the body that produced it has little influence on the hole's subsequent behaviour. The black hole is best thought of as a self-sustaining gravitational field governed by the internal non-linear dynamics of the general theory of relativity. These dynamics allow the asymmetries in the gravitational field of the hole to be carried away in the form of gravitational waves as the hole settles down into a stable configuration.

We have seen that a material object, once swallowed by a black hole, cannot escape. On the other hand, there are mechanisms whereby some of the energy content of the black hole can be extracted. One such mechanism involves the coalescing of two black holes. This process would be accompanied by the copious emission of gravitational waves, whose total energy should be a substantial fraction of the initial rest mass energy of the black holes. Another mechanism would be to allow a particle to fall into a region close to the event horizon of a rotating black hole. The particle splits into two particles in such a way that one falls into the hole and the other escapes back to infinity with *more* mass energy than the initial particle had. In this way rotational energy of the black hole is transferred to the particle motions outside the hole. In the process the black hole loses mass and spin. In principle, this is an extremely efficient means of converting rest-mass to energy—much more efficient than nuclear fission or fusion! But there is an absolute limit even to this procedure. In the most extreme case the mass of the black hole might conceivably be reduced down to 0.707 of its original value by this sort of general procedure. In any case it is hard to envisage this process being effective in an actual astrophysical situation.

Let us now consider the situation inside the black hole, and the general relativistic implication of the existence of a space-time singularity. Since a 'singularity' means a region of breakdown of physical theory, we have the curious situation that, here, general relativity is predicting its own downfall! But perhaps we should not be too surprised at this; after all we are treating general relativity only in its capacity as a classical theory. When the curvature of space time becomes enormous, quantum effects must eventually play a dominant role. When the radius of space-time curvature becomes as small as, say (10^{-13}) cm (roughly the radius of an elementary particle) then the theory of particle physics as understood at present, must break down. If the radius of space-time curvature ever becomes as small as (10^{-33}) cm (and the implications of what we have said so far are that it will be that small somewhere inside a black hole—unless theory breaks down before this), then we cannot avoid having to apply quantum mechanics to the structure of space-time itself. At present there is no satisfactory theory for doing this. It should be emphasised that there is no reason to believe that a new theory is needed to deal with situations less extreme than these. Only *at* the singularities would we have trouble. And if Cosmic Censorship holds true, the absolute event horizon would prevent the effects of such new physics having any influence on the outside world.

A question which often puzzles people in connection with black holes goes roughly as follows: If the absolute event horizon is so effective in shielding the contents of the black hole from outside, how is it that the black hole can still exert a gravitational influence on other bodies? How can it be that the gravitational field of the collapsed body escapes the black hole even though no information or signal can get out? In fact it is not really accurate to say that the gravitational field 'escapes'. It would be more true to say that the gravitational field is basically that of the body *before* collapse. If subsequently the body is annihilated at the centre of the hole, the exterior gravitational field cannot cut off. It would be *this* which would require information to escape the hole, because the gravitational field would have to 'know' when the body producing it had disappeared. So the exterior field does not reflect any change which takes place in the interior. After the body has collapsed in, it is better to think of the black hole as a self-sustaining gravitational field in its own right. It has no further use for the body which originally built it!

Let us return, now, to the observational status of black holes. The most promising place to look for black holes is in binary systems, such as HDE 226868—Cygnus X-1. As I mentioned earlier there are a number of such systems under consideration, with Cygnus X-1 being presently the most convincing candidate. Apart from the nature of the X-ray emission, the identification of the invisible component as a black hole would come from the estimate of its mass, this mass being detected via its gravitational effect on the visible component of the binary system. When the mass turns out to be considerably too large for a white or black dwarf or a neutron star, then the case for it being a black hole would seem to be a very strong one.

There is another aspect to the role of black holes in observational astronomy. We may compare the situation with that which arose in the case of the neutron stars. For many years astronomers had attempted to detect them by searching for certain effects such as X-ray emission which had been predicted also to occur in connection with neutron stars. However when neutron stars were first detected it was via an effect which was totally unexpected and still not really satisfactorily explained, namely the emission of the rapid and regular sharp pulses of electro-magnetic radiation characteristic of a pulsar. It is quite possible that the detection of other black holes will be by some equally unexpected observational side-effect.

There is no shortage of unexplained phenomena in cosmology today that might conceivably be relevant, from the phenomenal energy output of quasars and radio galaxies, the apparently anomalous red-shifts in the spectra of some quasars and galaxies, the explosions in the centres of some galaxies and the discrepancies in the mass measurements of galaxies, to serious questions that seem to remain even about the spiral-arm structure of normal galaxies. Above all, there is the apparent observation by Joseph Weber of the University of Maryland, of gravitational waves emanating from the centre of our own galaxy. As Professor Martin Rees points out in Chapter 10, if these waves are emitted continuously and in all directions uniformly from the galactic centre, the energy they carry would correspond to the loss of mass from the galaxy of many thousands of solar masses per year. This is in gross conflict with other observations.

So far the theory of black holes has not led to a very convincing explanation of this worrying phenomenon, or any of the others, but these are early days yet. In the case of Weber's observations taken at their face value, the best hope would seem to be some mechanism whereby the waves are strongly beamed in the galactic plane. The sun and the solar system are unusually close to the galactic plane—they lie in it to an accuracy of one part in a thousand. With a beaming angle of something like 1/1,000 the observational conflicts would be removed. Attempts have been made to explain this kind of beaming on the basis of a fast rotating black hole at the centre of our galaxy, but so far they are not very convincing. An alternative to beaming would be the possibility that the production of the waves from the galactic centre is a comparatively short-lived phenomenon, lasting for considerably less than a one-hundredth part of the time that our galaxy has been in existence. (This would be a rather uneasy coincidence, since our appearance on this earth presumably need not have occurred just at the time of extreme gravitational activity at the centre of the galaxy.) Another possibility is that there might be some source much nearer to us lying approximately on the straight line containing us and the galactic centre. However, there are objections to this idea too. Of course there is also the possibility that the observations are in some way spurious and that a phenomenon other than gravitational waves is contriving to produce the effects that Weber observes. A number of other groups have now built detectors and it is to be hoped that this uncertainty will be cleared up in a few years.

Let us assume, for the purposes of argument at least, that Weber's results are substantiated; then the question arises as to how nature produces such waves. Although black holes have usually been suggested as the most likely objects which could be responsible, no plausible detailed explanation seems to be forthcoming. I feel that we should not close our minds to the possibility that some other type of gravitational catastrophe might be occurring at the centre of our galaxy. The question of naked singularities has been largely left aside; but as I have said earlier, there is no very convincing theoretical argument in favour of Cosmic Censorship. Could it be that naked singularities *are* in some way responsible? I think that one should not ignore this possibility, particularly since it offers some hope of producing the beaming necessary to fit in with the observations.

Let me present the following very hypothetical picture. Suppose that some form of large rapidly rotating mass at one time collected at the centre of our galaxy and that this mass approached a regime at which the effects of general relativity dominated over all more conventional forces. Suppose also (and this is the big assumption) that such rotating agglomerations have a tendency to settle into a configuration whose exterior field closely approximates a Kerr solution of the Einstein field equations. Recall that black holes do just this—but I am not now thinking of a black hole! If the rotation remains too great, the solution will not in fact describe a black hole. Instead the matter would contract down until a naked singularity is revealed in a ring around the equatorial region. This of course is very speculative but the model possesses one feature which is perhaps suggestive: the naked singularity would be visible only from the equatorial plane! (This follows from some work by Brandon Carter.) Thus any signal originating at or near the singularity would necessarily be beamed very closely in one plane. If we assume a tie-up between this object and the structure of the galaxy then it is not unreasonable to suppose that the beaming plane and the galactic plane coincide. Finally, since ultra-strong gravitational tidal fields would exist near the singularity one would expect that it would be gravitational waves which would be so beamed. Weber's results might conceivably be explained in this way.

It has often been argued that if naked singularities arise then this situation would be disastrous for physics. I do not share such feelings. True we have, as yet, no theory which can cope with space-time singularities. But I am an optimist. I believe that eventually such a theory will be found. In any case we have for some time been confronted by the profound theoretical problems of the big bang singularity, and a theory is needed to cope with this. Would it not be far more exciting if in addition there were other space-time singularities accessible to view *now* which could supply observational means of testing such a new theory? Perhaps then the mysteries of the initial creation could be more readily comprehended.

60

NAKED SINGULARITIES

Sixth Texas Symposium on Relativistic Astrophysics. Ed. Dennis J. Hegyi. *Annals of the New York Academy of Sciences*, Vol. 224, pp. 125–34. New York Academy of Sciences,1973

In this article I toyed with the idea that cosmic censorship (see Chapters 40 and 270) might be *false*, coming across an inequality that if violated could indeed disprove cosmic censorship. This inequality (especially in the spacelike context, rather than the null one presented here) has become known as the 'Penrose inequality', and is of interest to (pure) mathematicians.

Naked singularities. Sixth Texas Symposium on Relativistic Astrophysics. Ed. Dennis J. Hegyi.
Annals of the New York Academy of Sciences, Vol. 224, pp. 125–34

Reprinted from
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 224, Pages 125–134
December 14, 1973

NAKED SINGULARITIES

Roger Penrose

Department of Mathematics
Birkbeck College
London W.C.1., England

I first review what is now the "standard" picture of gravitational collapse t black hole—which I shall refer to as the *establishment* viewpoint. The picture is fact, a very good one, with a remarkable degree of internal consistency. But i worth emphasizing that, among other things, this picture does depend on a v big assumption, that of *cosmic censorship*.[1] In my own view, this assumption mi well be false in suitable circumstances. I believe that the consequences of possi violations of cosmic censorship are certainly worth considering seriously, parti larly in view of Weber's observations.

The essentials of the establishment viewpoint are depicted in FIGURE 1. I am supposing that a collapse has taken place to the extent that one of the criteria for the applicability of the Hawking-Penrose singularity theorem[2] occurs. Let us say that either there is a trapped surface or, as is the case depicted in FIGURE 1, there is a future light cone (namely that of *P*) with the property that sufficient matter crosses it to cause all the light rays to begin to reconverge somewhere to the future of *P*. With the normal assumption that Einstein's field equations with "reasonable" sources (i.e., subject to the energy condition) hold and that the solution is adequately general and without closed timelike curves, the existence of some form of space-time singularity is implied. But the theorem does *not* say that the singularity is necessarily hidden inside an absolute event horizon. It is here that we must invoke the *ad hoc* assumption of cosmic censorship in order that the establishment picture be obtained. In effect, this principle states that naked singularities do not develop out of an initially nonsingular state, but we must be careful in formulating a precise statement of this, and I shall return to some of the difficulties later. A naked singularity may be said to exist in a space-time M if there is a well-defined external infinity (that is, $\mathscr{I}^+$, if M is weakly asymptotically simple[3]) from which timelike curves may be drawn into the past that terminate on the singularity (FIGURE 2). I shall not make precise, here, the concept of "terminating on a singularity" but refer instead to Hawking's concept of *asymptotic predictability*.[4] This is a precise condition on a space-time M that may be taken as a statement that M accords with a form of cosmic censorship. Imposing such a condition, we deduce the existence of an absolute event horizon E, which encloses but does not intersect the singular region. This horizon E may be defined as the boundary of the past of the external infinity. The property of E that its surface area increases with time may now be readily derived. This property has been used (by Floyd and Penrose[5] and by Hawking[4]) to obtain significant inequalities on the behavior of black holes. Finally, to complete the establishment picture we invoke the Israel-Carter-Hawking conjecture/theorem,[6-8] which asserts that the space-time settles down—in some appropriate but as yet ill-defined sense—to become that of a Kerr black hole. Unlike cosmic censorship, however, there is some considerable mathematical support for this particular conjecture. For the purposes of my discussion, therefore, I shall not call this conjecture into

Present address: Mathematical Institute, Oxford, England.

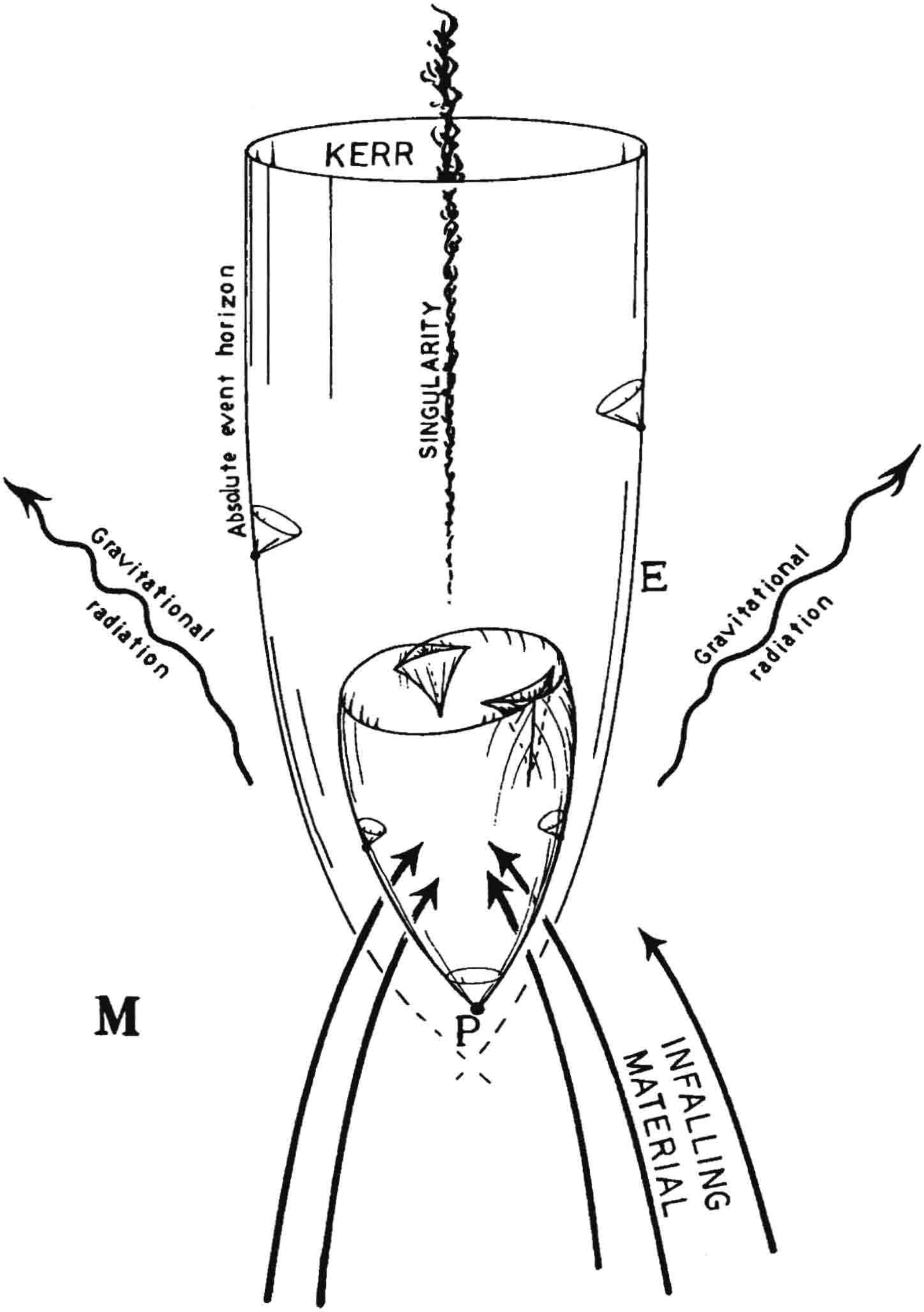

KERR
SINGULARITY
Absolute event horizon
E
Gravitational radiation
Gravitational radiation
M
P
INFALLING MATERIAL

FIGURE 1. Space-time diagram depicting collapse to a black hole according to the standard ("establishment") picture. (In all figures, time proceeds from the bottom of the page to the top.)

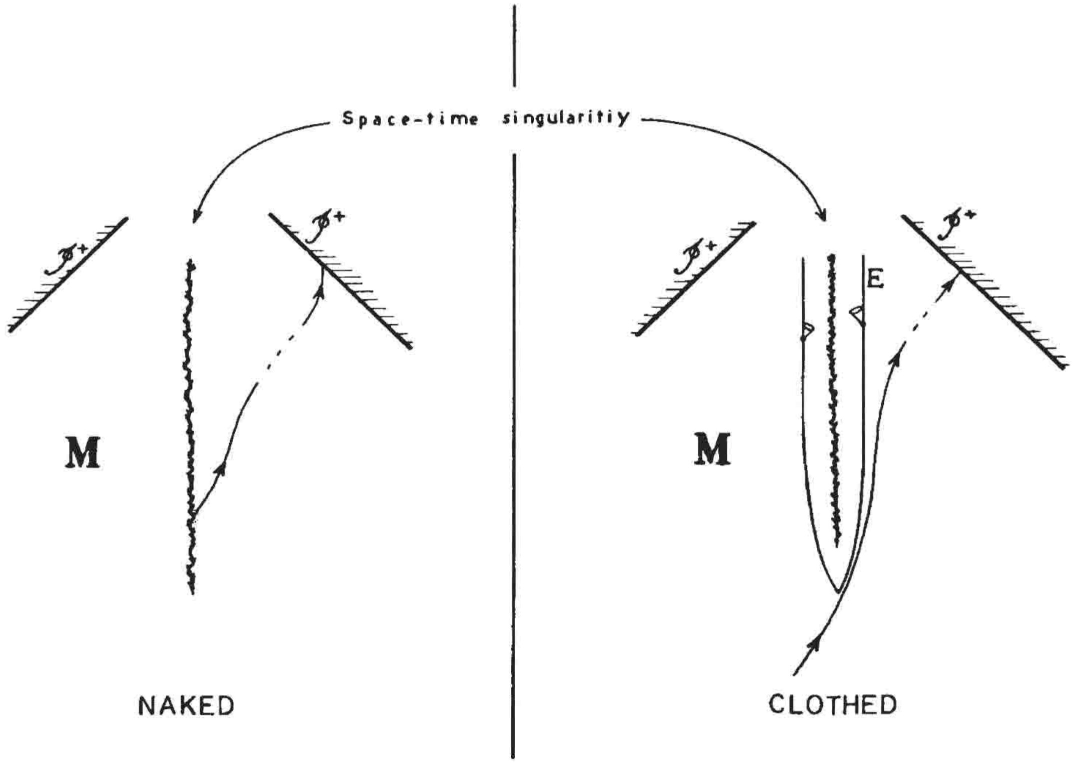


FIGURE 2. For asymptotically flat space-times, the difference between naked and clothed singularities is that in the first case the singularity lies to the past of $\mathscr{I}^+$ (future null infinity) but not in the second case.

question. But one should bear in mind that there is still a possible weakness in the establishment picture here too.

Let me describe an attempt that I made, about two years ago, to obtain a contradiction with the establishment point of view. I envisage a shell of photons ("null dust") that collapses into a region of flat space-time. (The situation is a kind of generalized time-reverse of Synge's "instantaneous transformation of a massive particle into radiation" [9].) I assume no symmetry either for the shape of the shell or for the distribution of energy density on the shell. The shell has merely to be convex—in order that no caustics occur in the remote past on the null hypersurface H, representing the history of the shell. This requirement (together with the fact that the shell is null) ensures that we can choose the space-time inside the shell to be flat (FIGURE 3).

Consider, now, an expanding null hypersurface N in the flat region, which intersects H in a (spacelike) two-surface S having the topology of a sphere. Extending N as a null hypersurface through H, we find that the divergence of N jumps down by an amount proportional to the intensity of energy of the shell at S. This focusing effect of H on N is smaller the farther out (i.e., further back into the past) that S is on H. This is because the energy intensity varies as r^{-2} back along the generators of H, whereas the divergence of N decreases as r^{-1}. Moving S inwards on H, we expect to find a location at which the divergence of N is exactly canceled so that the rays of N that extend into the future from S have zero divergence as they leave S. Let us suppose that $S = S_0$ can be so chosen. Then S_0 will be a marginally trapped surface.

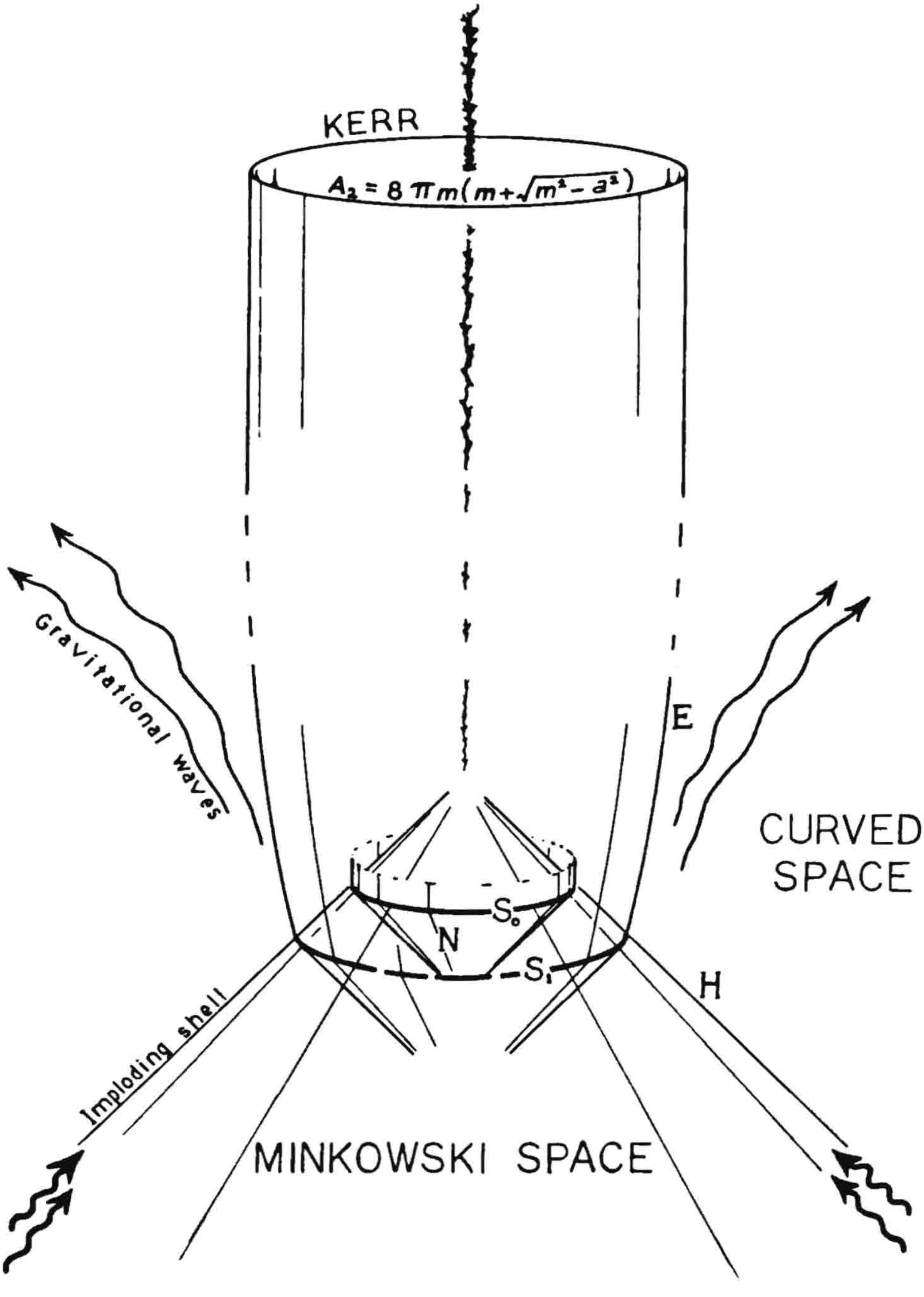

KERR
$A_2 = 8\pi m(m+\sqrt{m^2-a^2})$
Gravitational waves
E
CURVED SPACE
S_0
N
S_1
H
Imploding shell
MINKOWSKI SPACE

FIGURE 3. An attempt to derive a contradiction with the "establishment" viewpoint. A shell of photons collapses into a flat space-time. The assumption that a Kerr black hole results implies the existence of a certain inequality on the initial geometry.

If moved a little inwards along the generators of H, it becomes a trapped surface and the Hawking-Penrose theorem applies. Assuming cosmic censorship, we deduce the existence of an absolute-event horizon E intersecting H in a two-surface S_1, the surface S_1 lying farther out along the generators of H than S_0. Furthermore, the total surface area A_1 of S_1 is not less than the area A_0 of S_0. Moreover, the total surface area of cross-section of E increases from that of S_1 as we proceed into the future until, according to the establishment viewpoint, it approaches the final value

$$A_2 = 8\pi m(m + \sqrt{m^2 - a^2})$$

for the event horizon of the Kerr metric for mass, m, and angular momentum, am. (Units are chosen so that $G = c = 1$.) This in turn is less than or equal to the Schwarzschild value

$$A_3 = 16\pi m^2$$

for the same mass. This mass, m, cannot exceed the original total mass, m_0, of the collapsing shell, assuming that no further matter or radiation falls in from infinity. In general, we would expect the strict inequality $m < m_0$, since outgoing gravitational radiation would carry mass-energy away as the final black hole is formed.

We have, therefore, a string of inequalities

$$A_0 \leqslant A_1 \leqslant A_2 \leqslant A_3 \leqslant 16\pi m_0^2$$

implied by assumptions that go to make up the establishment picture of collapse to a black hole. Furthermore, A_0 and m_0 are quantities that can be related to each other and refer to the geometry of the original setup in flat space. If we could arrange such a collapsing shell to have $A_0 > 16\pi m_0^2$, then we should have a contradiction with the establishment picture. The most reasonable inference about the future history of the space-time resulting from such a collapse would then be that a naked singularity develops.

The initial setup is characterized simply by the two-surface S_0 embedded in the initial Minkowski space-time. A_0 is simply the area of S_0, while m_0 can be calculated from the geometry of the two null hypersurfaces, H and N, in Minkowski space, which intersect in S_0. The divergence of N at S measures the intensity of the energy of the shell at S_0 (since this intensity is exactly sufficient to reduce the divergence to zero). Scaling this intensity down by the inverse square of the appropriate luminosity parameter on H, we obtain the initial intensity of the shell at infinity. Integrating this over the sphere at infinity, we obtain m_0. For the case when the shell is spherical in shape (i.e., when H is a light cone) we may obtain the establishment inequality $A_0 \leqslant 16\pi m_0^2$ in the form of a simple mathematical relation

$$4\pi \oint r^2\Omega \leqslant \left(\oint r(1 - \nabla^2 \log r)\Omega\right)^2$$

where r is a smooth positive function defined on the unit sphere (with surface element $\Omega = \sin\theta\, d\theta \wedge d\varphi$ and covariant Laplacian ∇^2) the quantity r measuring the distance of S_0 from the time axis through the vertex of the cone H. (It is curious that exactly the same inequality arises for an analogous situation in which incoming gravitational waves, whose form is that of a curtailed Robinson-Trautman solution, implode to form a black hole.) We require the condition

$$\nabla^2 \log r < 1$$

to ensure positivity of the mass intensity.

Another special case of special interest is given when S lies in a spacelike hyperplane. Gibbons has pointed out that the inequality $A_0 \leqslant 16\pi m_0^2$ now becomes Minkowski's inequality

$$4\pi \oint \omega \leqslant \left(\oint K\omega\right)^2$$

for a closed convex surface in Euclidean three-space, with surface area element ω and extrinsic curvature K. Since this inequality is known to be valid, we see that no contradiction with the establishment picture is obtained for this situation. Gibbons has also shown that the general inequality $A_0 \leqslant 16\pi m_0^2$ follows from Minkowski's inequality by projection into a Euclidean three-space orthogonal to the time axis. However, Minkowski's inequality is not always true for surfaces that are not convex. In the general situation this projected surface need not be convex. Thus, it is still just possible that the inequality $A_0 \leqslant 16\pi m_0^2$ might be violated, but it could be violated only for a rather exotic initial setup.

The implications of all this are unfortunately rather inconclusive. I think it can be said that *some* support is offered to the establishment picture by Gibbon's work, at least if the collapse is of a kind that does not differ a great deal from that of spherical symmetry. But it might be argued that an exotic initial setup would in any case be needed if naked singularities were to be produced. It should be emphasized, also, that even if a proof of the inequality $A_0 \leqslant 16\pi m_0^2$ in *all* relevant circumstances were to hand, this would be a very long way off from establishing the principle of cosmic censorship. It is only that the *negation* of this inequality would, in effect, establish the *reverse*. It might be, for example, that a collapse to a naked singularity could occur without a trapped surface's developing at any stage. Finally, it may be remarked that even if no contradiction with the establishment picture can be obtained using the above argument, the inequalities obtained do have some physical implications. If we turn the argument around and *assume* the establishment viewpoint to be correct, we obtain inequalities limiting the amount of gravitational radiation that can be emitted for various configurations of imploding shells of radiation.

One aspect of the naked-singularity problem is the question of the stability of black holes. One may conceive of a situation in which a "black hole" is present at first but where the "event horizon" is unstable, the curvature in its neighborhood increasing to infinity with time. Although much significant work has been done on this subject e.g. refs. 10–12, one still cannot say with certainty whether the Kerr metric is stable. One possibility that has been suggested[13] is that for some range of values $k < a/m < 1$, with k some positive number less than one, it might be that the Kerr solution is unstable, although stable for $0 \leqslant a/m < k$. If an instability of this sort occurs, it might not, of course, be the kind that leads to a naked singularity. It could simply be that any excess angular momentum gets radiated away in gravitational waves.

I think I would be inclined to guess that the Kerr metrics for *all* $0 \leqslant a/m < 1$ *are stable* against small enough disturbances. But stability against large disturbances is much more problematical. The question has often been raised as to whether it might be possible to add high enough angular momentum material to a Kerr black hole to increase its a/m value to greater than unity. If this could be done, then a naked singularity would be the result, since for $a > m$ the Kerr metrics represent naked singularities and not black holes. However, the nearer that a approaches to m, the "harder" it is for large-angular-momentum prograde material to be swallowed by the black hole.[14,15] The principle of area increase may be invoked[5] to obtain a limit for the ratio of the increment of angular momentum $\delta(am)$ to the increment of mass δm in the form $\delta\{8\pi m(m + \sqrt{m^2 - a^2})\} > 0$. But since the area principle is based on cosmic censorship, we need to check this formula with an alternative means of derivation. In fact, such a derivation can be obtained, in the case of test fields, from the fact that the contraction $T_{ab}k^a$ of a Killing vector k^a with the test energy-momentum

$$T_{ab} \text{ (with } T_{ab} = T_{ba}, \nabla^a T_{ab} = 0)$$

gives a quantity subject to a global conservation law. Letting k^a be $(\partial/\partial t)^a$ and $(\partial/\partial\varphi)^a$ respectively, we obtain an exact conservation of test energy and test angular momentum about the axis. The area increase inequality now comes down to the fact that the flux $k^a T_{ab}\, dS^b$ across the event horizon, with surface element dS^b (taken future-null) must be positive—an implication independent of any cosmic censorship assumption. Because of the global conservation laws, this means that the asymptotically measured increment in test energy and test angular momentum must be subject to the same inequality. Thus, the increment in the total mass and angular momentum of the black hole should be subject to this inequality—so surface area increase may be reasonably inferred independently of cosmic censorship, with the implication that a gradual "spinning up" of a Kerr solution to values $a/m > 1$ cannot be achieved. (It may be remarked, however, that gravitational perturbations could conceivably present a different situation,[16] since no energy-momentum tensor exists for gravity.)

These considerations have very little to say about large perturbations, however. We might, for example, envisage two comparable black holes spiraling into one another. Have we any reason, other than wishful thinking, to believe that a black hole will be formed, rather than a naked singularity? Very little, I feel; it is really a completely open question. In fact, one question that should be asked is whether there is any *observational* evidence either for or against naked singularities occurring in nature. There are certainly enough odd things seen in astronomy that one is tempted to speculate about naked singularities playing a role. But perhaps the question should be asked the other way: is it possible to explain these observations *without* invoking naked singularities? The most glaring place where the answer might appear to be "no" is in Weber's results. If we take his conclusions at their face value, then we are led to ask how it might be possible to beam gravitational radiation very accurately in the galactic plane. Attempts, such as Misner's, to explain beaming of this kind in terms of black holes are certainly worthwhile, but it is very difficult to see how sufficient beaming can be obtained. I feel that it is also worth exploring the possibility that naked singularities could achieve the required effect.[17] Remarkably, the Kerr metrics with $a/m > 1$ do exhibit such a beaming. I shall consider a very hypothetical model that exploits this fact.

Suppose we have a rotating and slowly collapsing body, or collection of bodies, whose exterior field resembles a Kerr metric with $a > m$. I am going to suppose that, for some reason, when general-relativistic effects begin to dominate over all others, the Kerr configurations become increasingly preferred as approximate external fields. This is at the moment quite unjustified, since no event horizon is present and there is no other physical reason known for preferring Kerr fields. Nevertheless, it does not seem to me to be a *totally* unreasonable assumption in view of other special mathematical properties possessed by the Kerr field. Then, as the body shrinks, the external field approaches that of Kerr with $a > m$ ever more closely. The space-time curvature would become very large at the equatorial region. In the limit we could envisage the body attaining a disk-like form, the space-time curvature at its edge becoming virtually infinite—a naked singularity! Where curvatures are so large we would expect any disturbance to result in the emission of gravitational waves. But Carter[18] showed that signals (i.e., null geodesics) from the neighborhood of the singular ring can escape to infinity only in directions very close to the equatorial plane—the singularity itself being "visible" only from the equatorial plane. Thus, we have the potentiality for a very strong beaming effect for the gravitational waves—a beaming that could, in principle, be sufficient to remove the mass-loss conflict in Weber's observations.

Of course this is a highly speculative picture. However, I do not feel that it can be rejected out of hand. The existence of closed timelike curves in the Kerr solutions could have been a worrying feature, but the presence of the disk of rotating matter removes these. The main queries are whether an $a > m$ collapse would take place at all, and whether Kerr metrics are likely to arise in such situations if it does. It is, of course, possible that the beaming effect occurs with other naked singularities so that the restriction to Kerr might not be necessary. Unfortunately, very little is known about this. It is therefore worthwhile to point out some of the very remarkable properties that the Kerr metric actually does possess independently of the $a{:}m$ ratio, since they may possibly indicate a preference that nature could have for these metrics in highly general-relativistic situations.

In the first place the metrics are asymptotically flat, with two commuting Killing vectors (stationary and axi-symmetric). The Weyl tensor is type {22}, which implies the existence of a symmetric "Killing spinor," κ_{AB}, satisfying

$$\nabla_{(A}{}^{A'}\kappa_{BC)} = 0;$$

this means that the complex quantity $\pi^A\pi^B\kappa_{AB}$ is conserved along null geodesics, π^A being a parallelly propagated spinor pointing along the null geodesic. The conservation of $\arg(\pi^A\pi^B\kappa_{AB})$ implies that the propagation of the polarization plane along the null geodesic can be calculated directly without integration.[19] The integration of the geodesic equations (null or non-null) can reduced be to quadratures (as Carter first showed[18]). This fact can be derived from the existence of a symmetric Killing tensor, K_{ab}

$$\nabla_{(a}K_{bc)} = 0$$

in addition to the two Killing vectors.[19,20] R. Floyd[21] has shown that K_{ab} can be obtained from a different type of "Killing tensor," a skew tensor H_{ab} satisfying

$$\nabla_a H_{bc} = \nabla_{[a}H_{bc]}$$

where

$$K_{ab} = H_{ac}H_b{}^c.$$

The existence of such H_{ab}, together with the asymptotic flatness and vacuum conditions, is actually sufficient to characterize the Kerr family of space-time metrics uniquely. We have, in fact,

$$H_{ab} = i\kappa_{AB}\epsilon_{A'B'} - i\epsilon_{AB}\bar{\kappa}_{A'B'}$$

and

$$\kappa_{AB} = \psi^{-1/3}o_{(A}\iota_{B)}$$

where the Weyl conformal spinor is

$$\Psi_{ABCD} = \psi o_{(A}o_B\iota_C\iota_{D)},$$

with $o_A\iota^A = 1$. Floyd points out that the vector

$$F_a = H_{ab}P^b$$

is parallelly propagated along geodesics (tangent vector P^a) and conserved on collisions. The squared length F_aF^a is "Carter's fourth integral," given by $K_{ab}P^aP^b$, and is constant along geodesics. In some sense F_a seems to be giving a sort of measure of angular momentum that retains a measure of conservation about *any* axis, not just the rotation axis. The existence of such a quantity (together with asymptotic flatness and vacuum) occurs only for the Kerr metrics. But the physical implications of this fact are not yet clear.

To return to the question of cosmic censorship in a more general context, it must be said that still practically nothing is known about this. It is not even clear how one could formulate the principle in plausible but precise terms. Many solutions of Einstein's equations are known that do possess naked singularities (cf. the recent solution of Tomimatsu and Sato,[22] for example, or the Curzon[23] metric). It might even be possible to produce exact solutions representing a collapse of some special matter distribution to such a singularity. But this would in itself tell us rather little. We would have to know whether such behavior was "stable" in the sense of being qualitatively unchanged when the initial conditions are perturbed in some small but finite way.

As a postscript to this paper, I should also mention that some interesting solutions representing collapsing dust and possessing naked singularities have been given by Yodzis and coworkers.[24] These seem perhaps to be stable under perturbations of the initial conditions, but we need also to ask whether they are, in some appropriate sense, also stable under *perturbations of the equations of state*, e.g. what happens if pressure is introduced? These authors point out that Hagerdorn has suggested that there should be a maximum possible pressure to density ratio, so that when the density dominates this value, the pressures should become insignificant. Further work along such lines would seem to be of great relevance to the naked-singularity problem.

It is sometimes said that if naked singularities do occur, then this would be disastrous for physics. I do not share this view. We already have the example of the big-bang singularity in the remote past, which seems not to be avoidable. The "disaster" to physics occurred right at the beginning. Surely the presence of naked singularities arising occasionally in collapse under much more "controlled" circumstances would be the very reverse of a disaster. The effects of such singular occurrences could then be accessible to observation *now*. Theories of singularities would be open to observational test. The initial mystery of creation, therefore, would no longer be able to hide in the obscurity afforded by its supposed uniqueness.

References

1. Penrose, R. 1969. Rivista del Nuovo Cimento. Serie I, **1,** Numero Speciale: 252.
2. Hawking, S. W. & R. Penrose. 1970. Proc. Roy. Soc. (Lond.) A **314:** 529.
3. Penrose, R. 1968. *In* Battelle Rencontres C. M. DeWitt & J. A. Wheeler, Eds. W. A. Benjamin, Inc. New York, N.Y.
4. Hawking, S. W. 1972. Commun. Math. Phys. **25:** 152.
5. Penrose, R. & R. M. Floyd. 1971. Nature Phys. Sci. **229:** 177.
6. Israel, W. 1967. Phys. Rev. **164:** 1776.
7. Carter, B. 1971. Phys. Rev. Lett. **78:** 331.
8. Hawking, S. W. & G. F. R. Ellis. 1973. The Large-scale Structure of Space-Time (Cambridge Univ. Press.) Cambridge, England.
9. Synge, J. L. 1957. Proc. Roy. Irish Acad. A **59:** 1.
10. Doroshkevitch, A. G., Ya. B. Zel'dovich & I. D. Novikov. 1965. Zh Eksperim. i Teor. Fiz. **49:** 170; 1965. Soviet Phys. JET P **22.**
11. Vishveshwara, C. V. 1970. Phys. Rev. D **1:** 2870.
12. Price, R. H. 1972. Phys. Rev. D. **5:** 2439.

13. SMARR, L. 1972. Surface geometry of charged rotating black holes. Preprint.
14. BARDEEN, J. M. 1969. Nature **224:** 1263.
15. WALD, R. 1972. (Preprint).
16. PRESS, W. H. & S. TEUKOLSKI. 1972. Preprint.
17. PENROSE, R. 1972. Nature **236:** 377.
18. CARTER, B. 1968. Phys. Rev. **174:** 1559.
19. WALKER, M. & R. PENROSE. 1970. Commun. Math. Phys. **18:** 265.
20. HUGHSTON, L. P., R. PENROSE, P. SOMMERS & M. WALKER. 1972. Commun. Math. Phys. **27:** 303.
21. FLOYD, R. 1973. The Dynamics of Kerr Fields. (Ph.D. Thesis, London Univ. London, England.
22. TOMIMATSU, A & H. SATO. 1972. Phys. Rev. Lett. **29:** 1344.
23. CURZON, H. 1924. Proc. Lond. Math. Soc. **23:** 1344.
24. YODZIS, P., H.-J. SEIFERT & H. MÜLLER ZUM HAGEN. 1973. On the occurrence of naked singularities of general relativity. Preprint.

61

RELATIVISTIC SYMMETRY GROUPS

Group theory in non-linear problems (Lectures presented at the NATO Advanced Study Institute on mathematical physics, Istanbul, Turkey, August 7–18, 1972). Ed. A. O. Barut, pp. 1–58. Riedel Publishing Company, Dordrecht, 1974

648 Relativistic symmetry groups. *Group theory in non-linear problems* (Lectures presented at the NATO Advanced Study Institute on mathematical physics, Istanbul, Turkey, August 7–18, 1972). Ed. A. O. Barut, pp. 1–58

Roger Penrose
Department of Mathematics
Birkbeck College, London

1. ORTHOGONAL AND CONFORMAL GROUPS

A mathematical fact of very great significance for relativity theory is the existence of the familiar homomorphism* between the group SL(2,C) of complex unimodular (2 x 2) matrices and the Lorentz group O(1,3). This homomorphism

$$SL(2,C) \rightarrow O(1,3) \tag{1.1}$$

is a *local isomorphism* and maps SL(2,C) onto the identity-connected component of O(1,3), in an *essentially* (2-1) fashion. The term "essentially" here refers to the fact that SL(2,C) is connected. If A and B are the two elements of SL(2,C) which map to some given element Q of O(1,3) (actually B = -A), then A may be connected to B by some curve in SL(2,C). The image of this curve in O(1,3) is a closed curve κ through Q (topologically equivalent to a continuous rotation through 2π). Neither A nor B can be preferred as *the* SL(2,C) image of the Lorentz transformation Q. For as we pass from Q back to Q along the curve κ in O(1,3), the inverse image in

---

* A homomorphism between continuous groups is simply a continuous mapping from the first to the second which preserves the group operations. A local isomorphism is such a mapping which is 1-1 in the neighbourhood of the identity elements; then there is induced an isomorphism between the corresponding Lie algebras of infinitesimal group elements. The identity-connected component of a continuous group is its maximal connected subgroup (i.e. consisting of elements continuously connected with the identity element). For a discussion of the classical groups SL(2,C), O(1,3), etc. see references [1,2].

SL(2,C) must pass continuously from A to B, or else back from B to A. The ambiguity between A and B is therefore essential.

There is a higher-dimensional analogue of (1.1) which also has considerable importance for relativity theory, namely the homomorphism

$$SU(2,2) \rightarrow O(2,4) \qquad (1.2)$$

which is also a local isomorphism, and which similarly maps SU(2,2) onto the identity-connected component of O(2,4) in an essentially (2-1) fashion. The group SU(2,2) of unimodular pseudo-unitary (++--) (4 x 4)-matrices gives rise to the algebra of twistors, analogously to the way that SL(2,C) gives rise to the algebra of 2-component spinors. Twistors will be discussed in Section 4. The significance of the pseudo-orthogonal group O(2,4), for relativity theory, lies in its relation to the 15-parameter conformal group of Minkowski space-time. I shall denote this conformal group by C(1,3) and give its precise definition shortly. We have, in fact, a homomorphism

$$O(2,4) \rightarrow C(1,3) \qquad (1.3)$$

which is again a local isomorphism, mapping O(2,4) onto C(1,3) in an essentially (2-1) fashion. The homomorphism which is the composite of (1.2) with (1.3)

$$SU(2,2) \rightarrow C(1,3) \qquad (1.4)$$

is thus a local isomorphism which maps SU(2,2) onto the identity-connected component of C(1,3) in an essentially (4-1) fashion.

Similar to (1.3) is a homomorphism

$$O(1,3) \rightarrow C(2) \qquad (1.5)$$

where C(2) denotes a 6-parameter conformal group for the Euclidean plane analogous to C(1,3). (I shall be more precise shortly.) The homomorphism (1.5) is a local isomorphism which is (2-1), but it is not essentially (2-1). The map from the identity-connected component of O(1,3) onto the identity-connected component of C(2) is actually (1-1). The homomorphisms (1.3), (1.5) are part of a more general pattern of local isomorphisms:

$$O(p+1,\ q+1) \rightarrow C(p,q) \qquad (1.6)$$

The local isomorphisms (1.1) and (1.2) are, on the other hand, special features of the low dimensionalities involved. (We may note, in passing, there is the "non-relativistic" essentially (2-1) local isomorphism SU(2) → O(3), which is closely related to these, to quaternions and to non-relativistic spinors.) For the remainder of this section I shall be primarily concerned with (1.6) and its particular instances (1.3), (1.5). The special isomorphism (1.4) will play a basic role in Section 4.

First, consider (1.5). I have yet to define what I mean by the conformal group C(2). The orientation-preserving local conformal maps of the plane to itself may be conveniently represented by

$$\zeta \to \tilde{\zeta} = f(\zeta), \tag{1.7}$$

where f is a holomorphic (i.e. complex-analytic) function and where

$$\zeta = x + iy,$$

x and y being standard Cartesian coordinates for the plane. We have, for the line-element,

$$d\sigma^2 = dx^2 + dy^2 = d\zeta d\bar{\zeta} \tag{1.8}$$

so

$$d\sigma^2 \to d\tilde{\sigma}^2 = |f'(\zeta)|^2 \, d\sigma^2$$

illustrating the conformal nature of (1.7). Since f is arbitrary holomorphic, the local conformal maps of the Euclidean plane to itself constitute an infinite-parameter system. For a global map, we would require that both f and its inverse map be non-singular over the whole plane. This restricts f to be a linear function

$$f(\zeta) = \alpha\zeta + \beta$$

showing that the group of (orientation-preserving) conformal maps of the plane to itself is described by four real parameters (real and imaginary parts of α,β). These maps may be generated by the Euclidean motions ($|\alpha|=1$) and the dilations (α real, $\beta=0$).

This group is not what I mean by C(2), however. For that, we require to compactify the plane by the addition of a point at infinity. This is a standard procedure in complex variable theory, and is most graphically illustrated by means of a stereographic projection of the unit sphere S^2 to the plane (Figure 1). Let S^2 be given by the equation

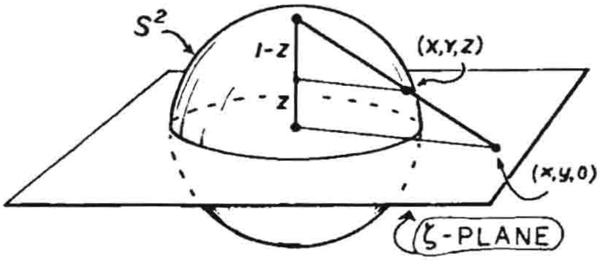


Fig. 1. The unit sphere S^2 is projected stereographically from the north pole (0,0,1) to the plane Z = 0, this plane being regarded as the Argand plane of the complex number $\zeta = x+iy$. The first formula (1.9) is readily obtained from the geometry of the picture.

$$X^2 + Y^2 + Z^2 = 1,$$

where X,Y,Z are standard Cartesian coordinates for Euclidean 3-space. We project the point (X,Y,Z) on S^2, from the north pole (0,0,1), into the point (x,y,0) on the equatorial plane Z=0, where

$$\zeta = x + iy = \frac{X + iY}{1 - Z}; \quad X + iY = \frac{2\zeta}{1 + \zeta\bar{\zeta}}, \quad Z = \frac{\zeta\bar{\zeta} - 1}{1 + \zeta\bar{\zeta}} \quad . \tag{1.9}$$

The complex number ζ, which previously labelled a point in the plane, may now be used to label the corresponding point (X,Y,Z) of the sphere S^2. We have one further point, however, namely the north pole of S^2 (corresponding to $\zeta = \infty$). The steroegraphic projection, defined by (1.9), of S^2 to the plane, is in fact a conformal map. One way of seeing this is to re-express ζ in terms of standard spherical polar coordinates θ,ϕ for the sphere

$$\zeta = e^{i\phi} \cot\frac{\theta}{2} \tag{1.10}$$

and then to observe that the metric $d\sigma^2$ for S^2 is given by

$$d\ell^2 = d\theta^2 + \sin^2\theta d\phi^2 = \frac{4d\zeta d\bar{\zeta}}{(1+\zeta\bar{\zeta})^2} = \frac{4d\sigma^2}{(1+\zeta\bar{\zeta})^2} \tag{1.11}$$

$d\sigma^2$ being the metric for the plane Z = 0, as in (1.8). Thus, S^2 with the north pole removed, is conformally identical with the Euclidean plane. The addition of the north pole provides the required *conformal compacitification* of the plane.

The group C(2) may now be defined as the group of all conformal maps of the compactified plane (i.e. of S^2) to itself. The connected component of the identity in C(2) consists of the orientation preserving conformal maps of S^2. These are given by (1.7) where f has the form

$$\zeta \to f(\zeta) = \frac{\alpha\zeta + \beta}{\gamma\zeta + \delta} \quad , \tag{1.12}$$

these being regular at $\zeta = \infty$. The three complex ratios $\alpha:\beta:\gamma:\delta$ serve to define f, so we have a *six*-real-parameter group. If desired, we can normalize $\alpha,\beta,\gamma,\delta$ by

$$\alpha\delta - \beta\gamma = 1. \tag{1.13}$$

Then the unimodular complex (2 x 2) matrix (i.e. SL(2,C) element)

$$\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \tag{1.14}$$

may be used to represent our transformation (with only an overall sign ambiguity for (1.14)) the composition of two transformations (1.12) being now represented by matrix multiplication. (This can be used to derive the local isomorphism SL(2,C) → C(2).) The remaining (i.e. orientation-reversing) elements of C(2) are the compositions of (1.12) with complex conjugation: $\zeta \to \zeta$. The elements of C(2) are generated by the Euclidean motions, dilations, and the inversion

$$\zeta \to \tilde{\zeta} = \bar{\zeta}^{-1} \; . \tag{1.15}$$

(This inversion is disconnected from the identity.) The group C(2) is transitive over the whole of S^2, showing that the adjoined point at infinity is (conformally) on an equal footing with all the other points. The inversion (1.15) actually interchanges the origin $\zeta = 0$ with the point at infinity $\zeta = \infty$ (i.e. it interchanges the north and south poles of S^2, being a reflection in the equatorial plane).

To establish the relation of C(2) with O(1,3), consider Minkowski 4-space with standard coordinates T,X,Y,Z, the metric being given by

$$ds^2 = dT^2 - dX^2 - dY^2 - dZ^2 \; .$$

I shall be concerned primarily with the null cone $\mathcal{N}$ of the origin, its equation being

$$T^2 - X^2 - Y^2 - Z^2 = 0. \tag{1.16}$$

The *generators* of $\mathcal{N}$ are the null rays through the origin, given by

$$T:X:Y:Z = \text{const.}$$

with T,X,Y,Z satisfying (1.16). Let us consider S^2 to be the section of $\mathcal{N}$ by the spacelike 3-plane T = 1. Then there is a (1-1)-correspondence between the generators of $\mathcal{N}$ and the points of S^2 (namely that given by the intersections of the generators with T = 1). We may regard S^2 as a realization of the *space of generators* of $\mathcal{N}$. But equally well we could have used any other cross-section $\hat{S}^2$ of $\mathcal{N}$ to represent this space. However, the important point to realize is that the map which carries any one such cross-section into another, with points on the *same* generator of $\mathcal{N}$ corresponding to one another, is a *conformal* map (Figure 2). Thus,

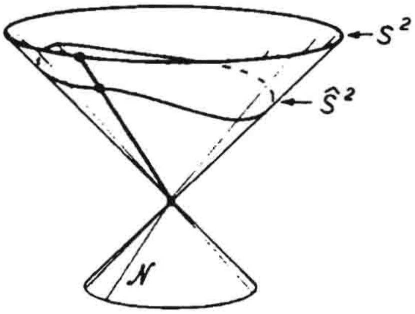


Fig. 2. The generators of the null cone $\mathcal{N}$ establish a 1-1 map between any two cross-sections of $\mathcal{N}$. This map is conformal, so the space of generators of $\mathcal{N}$ may itself be assigned a conformal structure, namely that of any one of these sections.

the conformal structure of S^2 - or, equivalently, of $\hat{S}^2$ - reflects an intrinsic conformal structure on the space of generators of N.

To see that this generator map is a conformal map we may re-express the metric on N in a form

$$ds^2 = -r^2\gamma_{\alpha\beta}dx^\alpha dx^\beta + 0.dr^2 \qquad (1.17)$$

where x^α and r are coordinates on N, the generators being given by the coordinate lines $x^\alpha = \text{const}$. The quantities $\gamma_{\alpha\beta}$ are to be independent of r. (There are clearly many ways of attaining the form (1.17); one is to use ordinary spherical polar coordinates for Minkowski space, giving $ds^2 = dT^2 - dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2)$, but with $T = r$ on N). A cross-section of N is given by specifying r to be some function of the x^α. It is obvious from the form of (1.17) that any two cross-sections give conformally related metrics, being mapped to one another by the generators of N (i.e. keeping x^α constant). It is clear, moreover, that many other cone-like null surfaces will share this property of N, provided their metrics can be put in the form (1.17).

Let us now consider the effect of an (active) Lorentz transformation L on our configuration. This will send N into itself and send generators of N into other generators of N. It will send the cross-section S^2 into some other cross-section (say $\hat{S}^2$) whose intrinsic metric is the same as that of S^2. Since the map from $\hat{S}^2$ back to S^2 along generators of N is conformal, it follows that L induces a transformation on the space of generators of N which amounts to a conformal map of S^2 to itself. This establishes a homomorphism $O(1,3) \rightarrow C(2)$. Since the dimensionality of each group is the same, namely six, the inverse image of each element of C(2) must be a discrete set of elements of O(1,3). From this it follows that the mapping is a local isomorphism. Indeed, it is intuitively clear that L, up to a space-time reflection in the origin, is defined by its effect on N. The homomorphism $O(1,3) \rightarrow C(2)$ is thus (2-1). But it is only inessentially (2-1). For restriction to orthochronous O(1,3) transformations would yield a global isomorphism between the two groups.

If, instead of the cross-section of N by $T = 1$, we consider the intersection of N with the null 3-plane $T = Z + 1$, we get a "parabolic" section E^2 whose intrinsic metric is $d\sigma^2 = -ds^2 = dX^2 + dY^2$. Thus, E^2 is intrinsically a Euclidean plane. It is not quite a cross-section of N because the one generator $T - Z = X = Y = 0$, which is parallel to the 3-plane, fails to give rise to a point on E^2. Nevertheless the relation between E^2 and S^2, for all other generators, is a conformal one. The generator parallel to the 3-plane simply corresponds to the point at infinity on E^2. In fact, the sterographic projection considered earlier may be established by examining 2-planes through this generator (see Figure 3). Such a 2-plane meets S^2, E^2, and the plane $Z = 0 = T - 1$ in corresponding points. The details are left as an exercise.

Let us next consider $O(2,4) \rightarrow C(1,3)$. This will also serve to

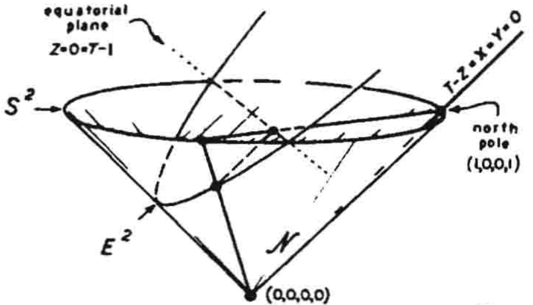


Fig. 3. The Euclidean plane E^2 may be imbedded as a "parabolic" section of N. Its conformal relation to the unit sphere S^2 is established via generators of N, or via 2-planes through the "parallel" generator $T - Z = X = Y = 0$.

illustrate the general case $O(p + 1, q + 1) \to C(p,q)$, which is basically no more complicated. In the first place, it will be necessary to define the group $C(1,3)$ (or, more generally, $C(p,q)$) precisely. As in the case of $C(2)$, we may regard $C(1,3)$ as the group of conformal self-transformations of the appropriate pseudo-Euclidean space (in this case, Minkowski 4-space) which has been compactified [3,4,5,6] in a suitable conformal way, by the addition of extra "points at inifinity". In the present context, the most rapid way of achieving this conformal compactification of Minkowski space is to go directly to the null cone of the origin in a six-dimensional pseudo-Euclidean (++----) space. Let us choose coordinates T, V, W, X, Y, Z for this space, the metric being given by

$$ds^2 = dT^2 + dV^2 - dW^2 - dX^2 - dY^2 - dZ^2 \,, \tag{1.18}$$

the null cone K of the origin having the equation

$$T^2 + V^2 - W^2 - X^2 - Y^2 - Z^2 = 0 \,. \tag{1.19}$$

By analogy with E^2 above we can consider the intersection M^4 of K with the null hyperplane $V - W = 1$. The intrinsic metric of M^4 is

$$ds^2 = dT^2 - dX^2 - dY^2 - dZ^2 \,.$$

The coordinates T, X, Y, Z suffice for M^4 and are unrestricted in range. Thus, M^4 is intrinsically identical with Minkowski spacetime. In the 6-space, however, M^4 has the form of a "paraboloid", the remaining coordinates being defined in terms of T, X, Y, Z by

$$V = \tfrac{1}{2}\,(1 - T^2 + X^2 + Y^2 + Z^2) = W + 1 \,.$$

Every generator of K (set of points for which $T:V:W:X:Y:Z$ are constant and for which (1.19) is satisfied), except for those lying in the null hyperplane $V = W$, will meet M^4 in a unique point.

We may think of those generators of K which do lie in V + W as representing "points at infinity" for M^4. Thus, the set G of (unoriented) generators of K (being a compact topological space) may be interpreted as a "compactification" of M^4, those members of K which do not lie in V = W being in (1-1)-correspondence with M^4, and those which do lie in V = W supplying the extra points necessary for the compactification. Now, as with the case of N above, the metric of K can (in many ways) be put into the form (1.17). (For example, we may select any variable, say W, and re-express (1.18) as $ds^2 = W^2\{d(T/W)^2 + d(V/W)^2 - d(X/W)^2 - d(Y/W)^2 - d(Z/W)^2\}$ using (1.19); and then eliminate one of the redundant variables T/W, ---, Z/W by expressing it in terms of the others, again using (1.19).) Thus, as with N, the generators of K establish conformal maps between any two (local) cross-sections of K. This gives G the structure of a conformal manifold, which we can identify as the conformal compactification of M^4. The group C(1,3) is now defined as the group of self-transformations of G which preserve its conformal structure. The definition of C(p,q) is similar.

There are certain differences arising in the case of C(1,3) from the earlier situation for C(2). In the first place, the infinite parameter set of "local" conformal transformations of the plane given by (1.7) has no analogue for dimension higher than two, and so does not occur here. On the other hand, the global conformal self-transformations of M^4 are like those for E^2. They may be generated by (pseudo-) Euclidean motions (here constituting the Poincaré group) and dilations. The 11-parameter (orthochronous) group thus arising is sometimes known as the causal group. In "compactifying" M^4 we must now adjoin not just a point but an entire null cone. One way of seeing this is to examine the "inversion" of M^4, which is the analogue of (1.15). This corresponds to the reflection W → -W in 6-space and can be expressed

$$(T, X, Y, Z) \rightarrow -\{T^2 - X^2 - Y^2 - Z^2\}^{-1} (T, X, Y, Z) \qquad (1.20)$$

in terms of Minkowski coordinates. The transformation (1.20) is not well-defined on the null cone $T^2 - X^2 - Y^2 - Z^2 = 0$, but maps this entire null cone to infinity. In effect, (1.20) "exchanges" the null cone of the origin with the null cone at infinity, and illustrates the fact that the elements at infinity must actually have the structure of a null cone.

As an analogue of the sphere S^2, we could consider the intersection of K with the hyperplane T = 1. This intersection has the structure of a de Sitter space. However, the de Sitter space does not represent the entire compact space G, there still being points at infinity (corresponding to the generators of K in T = 0). Similar remarks apply to any hyperplane section of K. In particular, the "anti-de Sitter space" , defined by the section of K with W = 1, also requires points at infinity to be added in order to become compact. Thus, none of these space-times is an adequate model for the compactified Minkowski space G. As an alternative, we can consider the intersection of K with the 5-sphere defined by

$$T^2 + V^2 + W^2 + X^2 + Y^2 + Z^2 = 2 .$$

This gives a compact space-time model $\mathcal{H}$ which (by (1.19)) is clearly the topological product of a 3-sphere in (W, X, Y, Z) - space

$$W^2 + X^2 + Y^2 + Z^2 = 1$$

with a circle (1-sphere) in (T, V) - space

$$T^2 + V^2 = 1 .$$

However, this is not quite a model of G, but a twofold covering of it. This is because each point of H is represented once as (T,V,W,X,Y,Z) and once as (-T,-V,-W,-X,-Y,-Z) on $\mathcal{H}$. The space-time $\mathcal{H}$ is connected, so the two-fold nature of the covering is "essential". The topology of G is S^1 x S^3, as is its twofold covering.

The pseudo-orthogonal group O(2,4) acts on the 6-space. Since K is invariant, the group also acts on K. Each O(2,4) transformation must induce a conformal map of G to itself, and a homomorphism O(2,4) → C(1,3) is thereby obtained. The group O(2,4) has 15 parameters; C(1,3) cannot have more parameters than 15 (this being the maximum for local conformal self-transformations of a 4-manifold [7]). The fact that the homomorphism is a local isomorphism is not hard to establish from this. The fact that it is (2-1) arises because O(2,4) transformations can reverse the directions of the generators of K, whereas G is the space of *unoriented* generators. The O(2,4) transformation which is a reflection in the origin represents the identity on G. This reflection is continuous with the identity because O(2) x O(4) is contained in O(2,4) and the reflection in the origin for each of O(2) and O(4) is continuous with the identity. The essentially (2-1) nature of the homomorphism O(2,4) → C(1,3) follows from this. A similar remark applies to each O(p + 1, q + 1) for which both p and q are odd. If both are even, the local isomorphism O(p + 1, q + 1) → C(p, q) is still (2-1) but inessentially so; if p + q is odd the local isomorphism is (1-1) and is therefore a global isomorphism.

A smooth map from one space-time to another may be characterized as conformal by the fact that it takes null cones into null cones, or that it takes null geodesics into null geodesics. It is of some interest, therefore, to observe that the null line and null cone structure of (compactified) Minkowski space is reflected in the linear structure of K. There are, in fact ∞^5 2-planes through the origin, in the 6-space, which are completely contained in K. These planes are *completely null* in the sense that the distance (as defined by (1.18)) between *any* two points on each of the planes is zero. The intersection of such a plane with M^4 is therefore a null straight line in M^4. The generators of the null cone at infinity for M^4 also arise from such 2-planes, in this case lying in V = W. The 5-plane V = W is a tangent 5-plane to *K*, which touches *K* all along the particular generator V - W = T = X = Y = Z = 0 corresponding to the vertex of the null cone at infinity for M^4. The intersection of V = W with *K* defines this null cone. All "finite" null cones on M^4 also arise as intersections with tangent 5-planes to *K* (i.e. null 5-planes through the origin of 6-space). A different approach to the geometry of compactified Minkowski space will be discussed in the next section.

2. ASYMPTOTICALLY SIMPLE SPACE-TIMES

The construction of a "conformal infinity" for Minkowski space can be approached from another point of view which lends itself more readily to a generalization to curved space-times. According to this alternative approach [6] we simply rescale the metric of the space-time *M*, replacing the original physical metric ds by a new "unphysical" metric $d\hat{s}$, which is conformally related to it

$$d\hat{s} = \Omega ds \ , \qquad (2.1)$$

Ω being a suitably smooth, everywhere positive function defined on *M*. The metric tensor* g_{ab} and its inverse g^{ab} are accordingly rescaled by

$$g_{ab} \to \hat{g}_{ab} = \Omega^2 g_{ab} \ , \ g^{ab} \to \hat{g}^{ab} = \Omega^{-2} g^{ab} \ . \qquad (2.2)$$

Provided that the asymptotic structure of *M* is suitable, and that Ω is chosen appropriately, it is possible to adjoin more points to the manifold in such a way that the "unphysical" metric $\hat{g}_{ab}$ extends smoothly to these new points. The function Ω can also be extended smoothly, but becomes zero at these extra points. This implies that the physical metric would have to be infinite at the new points, so it cannot be so extended. Thus the new points are, from the point of view of the physical metric, infinitely distant from their neighbours. Physically, they represent points "at infinity".

It should be emphasized that the conformal rescaling (2.1), (2.2) is not a conformal transformation (conformal mapping) of the kind considered in Section 1. Here the points of the manifold are not transformed. It is only that one metric on a fixed manifold is

---

* From now on I shall be adopting the "abstract index" conventions according to which the symbol "g_{ab}" actually stands for the metric tensor itself rather than its components in some coordinate system. This makes little difference at this stage, but it has the effect of simplifying some of the formulae arising when the spinor formalism is employed towards the end of this section and in Section 4. See ref. [8] for details.

replaced by another. Conformal rescalings on a given space-time form an (uninteresting) infinite parameter Abelian group, namely the multiplicative group of smooth positive functions on the manifold. (This group, as applied to Minkowski space, has no element in common with the 15-parameter conformal group - except, perhaps, for the identity element - and should not be confused with that conformal group.) It is sometimes convenient to accompany a conformal rescaling with a coordinate change. This is merely in order that the newly adjoined points be assigned finite coordinates. No transformation of the points of the space-time is involved. Coordinates are, in any case, just a matter of convenience. If desired, the new coordinates may be introduced first, before any conformal rescaling takes place. This serves to emphasize that the transformation involved (namely the conformal rescaling) has nothing to do with a change in coordinates. The calculations that one performs using this technique are, in any event, often of the invariant, basically coordinate-free type.

The utility of this technique derives, to a large extent, from the fact that many important physical concepts are actually invariant under conformal rescalings [6]. In particular, the concepts of causality and of a null geodesic (or light ray) are invariant; so also are the zero rest-mass free-field equations for each spin, and electromagnetic interactions. On the other hand, timelike (or spacelike) geodesics, fields of non-zero rest-mass, and gravitational interactions are not conformally invariant. I shall discuss some of these matters a little more fully in due course. The conformal technique is valuable particularly when radiation properties of zero rest-mass fields are to be discussed. The incoming and outgoing radiation fields may be defined precisely, for an asymptotically flat, curved space-time, in terms of the values of the fields at the adjoined "points at infinity". The conformal technique also affords a very convenient and "coordinate-free" definition of asymptotic flatness in general relativity. It becomes unnecessary to make statements about asymptotic properties of the space in terms of complicated limiting procedures. Invariance properties are readily discerned; complicated coordinate transformation properties are avoided. Finally, as we shall see in Section 3, the Bondi-Metzner-Sachs asymptotic symmetry group can be given a clear geometrical interpretation, and its relation to gravitational radiation readily understood.

Let us first see how the technique may be applied in the case of Minkowski space. The physical metric, in spherical polar coordinates, is

$$ds^2 = dt^2 - dr^2 - r^2(d\theta^2 + \sin^2\theta\, d\phi^2) \,. \tag{2.3}$$

For convenience, introduce a retarded time parameter $u = t - r$ and an advanced time parameter $v = t + r$ to obtain

$$ds^2 = du dv - \tfrac{1}{4}\,(v - u)^2\,(d\theta^2 + \sin^2\theta\, d\phi^2) \,. \tag{2.4}$$

There is much freedom in the choice of conformal factor Ω. However, for the type of space-time that we shall consider here ("asymptotically simple" space-times), our choice of Ω must be such that along any null geodesic it approaches zero, both in the past and in the future, like the reciprocal of an affine parameter λ on the geodesic (i.e. $\Omega\lambda \to$ const. as $\lambda \to \infty$, and $\Omega\lambda \to$ const. as $\lambda \to -\infty$, along the geodesic). Each u = const. hypersurface is a future light cone, generated by null geodesics (straight lines) given when θ and ϕ are also constant. The coordinate v serves as an affine parameter into the future on each of these null geodesics. Similarly, the coordinate u serves as an affine parameter into the past on these radial null geodesics. Thus, we shall require $\Omega v \to$ const. as $v \to \infty$ on $u, \theta, \phi =$ const. and $\Omega u \to$ const. as $u \to -\infty$ on $v, \theta, \phi =$ const. If we wish also to keep Ω smooth over the finite parts of the space-time, then the choice

$$\Omega = (1 + u^2)^{-1/2} (1 + v^2)^{-1/2}$$

suggests itself, so

$$d\hat{s}^2 = \Omega^2 ds^2 = \frac{dudv}{(1+u^2)(1+v^2)} - \frac{(v-u)^2}{4(1+u^2)(1+v^2)} (d\theta^2 + \sin^2\theta d\phi^2).$$

Many other choices of Ω are equally possible, but this one is especially convenient, as we shall see shortly.

In order that our "points at infinity" may be assigned finite coordinates, we can replace u and v by p and q, where

$$u = \tan p, \quad v = \tan q .$$

Then

$$d\hat{s}^2 = dpdq - \tfrac{1}{4} \sin^2(q-p) \{d\theta^2 + \sin^2\theta d\phi^2\} . \tag{2.5}$$

The range of the variables p, q is as indicated in Figure 4. The vertical line q − p = 0 represents the spatial origin (r = 0) and is just a coordinate singularity. The space-time is, of course,

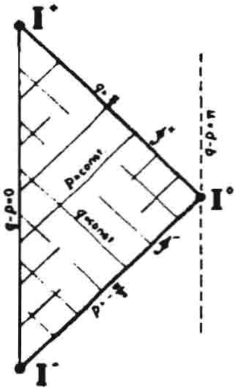


Fig. 4. The range of the coordinates p, q for Minkowski space.

non-singular on this line. The sloping lines $p = -\pi/2$ $(-\pi/2 < q < \pi/2)$ and $1 = \pi/2$ $(-\pi/2 < p < \pi/2)$ represent (null) infinity (denoted $\mathcal{I}^-$ and $\mathcal{I}^+$, respectively) for Minkowski space (since they correspond to $u = -\infty$ and to $v = \infty$). However, the metric (2.5) is evidently perfectly regular on these regions. Indeed, the space-time and its metric $d\hat{s}$ can clearly be extended beyond these regions in a non-singular fashion. The vertical line $q - p = \pi$ is again a coordinate singularity - of precisely the same type as that at $q - p = 0$. The entire vertical strip $0 \leq q - p \leq \pi$ may be used to define a space-time $\mathcal{E}$ whose global structure is that of the product of a space-like 3-sphere with an infinite timelike line (an "Einstein static universe"). To see this, we choose new coordinates

$$T = \tfrac{1}{2}(p + q), \quad \rho = q - p$$

and we have

$$d\hat{s} = dT^2 - \tfrac{1}{4}\{d\rho^2 + \sin^2\rho(d\theta^2 + \sin^2\theta\, d\phi^2)\} .$$

The part in curly brackets represents the metric of a unit 3-sphere.

The portion of $\mathcal{E}$ which is conformal to the original Minkowski space may be described as that lying between the light cones of two points I^- and I^+. The point I^- is given by $p = q = -\pi/2$, and I^+ by $p = q = \pi/2$. This portion "wraps around" $\mathcal{E}$ to meet at the "back" in the single point I° (given by $q = -p = \pi/2$). (Note that $\sin^2(q-p) = 0$ at I°, indicating that I° should, in fact, be regarded as a single point, rather than a 2-sphere.) The situation is illustrated in Figure 5 in the two-dimensional case. Minkowski 2-space is conformal to the interior of a square (represented as

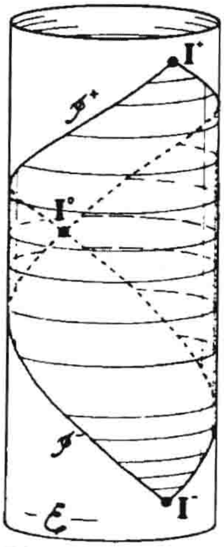


Fig. 5. Minkowski space is conformally identical to a portion (shown shaded) of an "Einstein static universe". The boundary of this portion represents the conformal infinity of Minkowski space.

tipped at 45°). This square wraps around the cylinder which is the two-dimensional version of the Einstein static universe. In higher dimensions the situation is similar. Near the point I^-, the relevant region lies in the *interior* of the *future* light cone of I^-. This light cone (i.e. the set of null geodesics directed into the future from I^-) is focussed around the back of the Einstein universe to a single point I° (which is spatially the antipode of I^-). Near I° the relevant ("Minkowski") region lies in *spacelike* directions from I°. The future light cone of I° is focussed back again to a single point I^+, whose spatial location corresponds to that of I^-. Near I^+, the relevant region lies in the *interior* of the *past* light cone of I^+ (see Figure 6).

The null geodesic segments which connect I^- to I° sweep out the portion of the boundary of the Minkowski space region that has been denoted* $\mathcal{I}^-$. Similarly the null geodesic segments from I° to I^+ sweep out $\mathcal{I}^+$. The points I^-, I°, I^+ themselves are considered not to belong to $\mathcal{I}^-$ or to $\mathcal{I}^+$. Physically, we interpret I^- as representing past temporal infinity, $\mathcal{I}^-$ as past null infinity, I° as spatial infinity, $\mathcal{I}^+$ as future null infinity, and I^+ as future temporal infinity. The reason for this terminology is made clear if we examine the behaviour of straight lines in Minkowski space (straight, that is, according to the Minkowski metric ds). A timelike straight line acquires a past end-point I^- and a future end-point I^+. A null straight line acquires a past end-point on $\mathcal{I}^-$ and a future end-point on $\mathcal{I}^+$. A spacelike straight line becomes a closed curve through I° when the point I° is adjoined. (The detailed verification of these facts is left as an exercise.) Since null geodesics remain null geodesics after conformal rescaling, we have

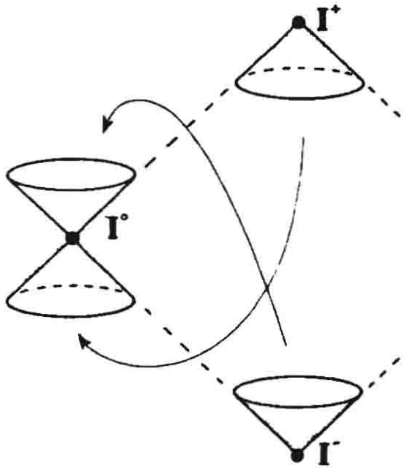


Fig. 6. The regions shaded, near each of I^-, I°, I^+, represent parts of the Minkowski space. If I^-, I°, and I^+ are identified as a single point, these three regions fit together.

---

* In order to distinguish verbally between $I^\pm$ and $\mathcal{I}^\pm$ in a concise fashion, it is convenient to pronounce $\mathcal{I}$ as "scri" - a contraction of "script I". See figures for the usual notation for $\mathcal{I}$.

the fact that the null straight lines become null geodesics according to the $d\hat{s}$ metric (but the timelike or spacelike straight lines are not, in general, geodesics with respect to $d\hat{s}$).

When we come to consider asymptotically flat space-times shortly we shall see that much of the above discussion will also apply to asymptotically flat space-times. However, one property which is very specific to the Minkowski space model is the fact that every null geodesic which originates at some point a^- on $\mathcal{I}^-$ will pass through the same point a^+ on $\mathcal{I}^+$ (see Figure 7). This property may seem surprising at first, but it becomes clear when we realize that the future light cone of a point of $\mathcal{I}^-$ is simply a null hyperplane in the Minkowski space. (It is the limit of a light cone when the vertex recedes into the past along a null straight line.) Similarly the past light cone of any point on $\mathcal{I}^+$ is also a null hyperplane. So a null hyperplane will acquire a past "vertex" on $\mathcal{I}^-$ (say a^-) and a corresponding future "vertex" (say a^+) on $\mathcal{I}^+$. We can also see this in terms of Einstein universe model $\mathcal{E}$. The future light cone of a^- will be focussed at a point which is spatially antipodal to a^-. This will be the point a^+.

Having this natural association between points of $\mathcal{I}^-$ and points of $\mathcal{I}^+$, for Minkowski space, it is reasonable to perform an identification between $\mathcal{I}^-$ and $\mathcal{I}^+$, the point a^- being identified with a^+. If we do this, then for the sake of continuity we should also identify I^- with I°, and I° with I^+. The three points $I^\pm$, I° thus become one - which we label I. We see from Figure 6 that the Minkowski regions fit neatly together at the point I, so that I becomes simply a normal interior point of the identified manifold. In fact, the manifold that we have constructed by performing these identifications is simply the "compactified Minkowski space" that we obtained in another way in Section 1. Since a^- is identified with a^+, each null geodesic becomes closed, with the topology of a circle S^1. Regarding only the conformal structure, every point of this space is on an equal footing with every other point, whether it be I itself, on the light cone $\mathcal{I}^- = \mathcal{I}^+$ of I, or in the finite portion of the Minkowski space.

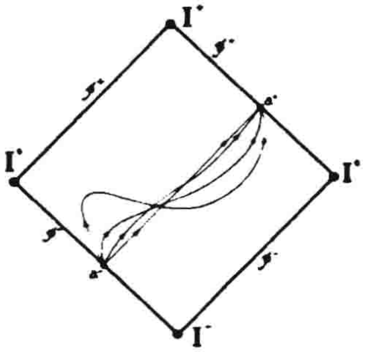


Fig. 7. Conformal infinity for Minkowski space has the special property that every null geodesic in the space which originates at $a^- \in \mathcal{I}^-$ must terminate at the same point $a^+ \in \mathcal{I}^+$.

When discussing Minkowski space it is sometimes convenient to perform the above identifications; sometimes it is more convenient not to and to leave I^- and I^+ as distinct boundary hypersurfaces ("conformal infinity"). In the case of curved asymptotically flat space-times, however, only the second course of action is reasonable. There are two reasons for this. In the first place there appears to be, in general, no natural association of a point a^- on I^- with some unique point a^+ on I^+. For example, the null geodesics from a^- will not focus cleanly at a point of a^+, but will tend to cross over one another (and to encounter "caustics") before I^+ is reached. However, the situation is worse than this. Suppose some suitable "canonical" scheme for making identifications is found. It turns out that even in the simplest cases (e.g. the Schwarzschild solution) the performing of any identification of I^- with I^+ will result in singularities in the metric $d\hat{s}$ along I, whatever the choice of Ω (although $d\hat{s}$ can, in some cases, be made C^2 - but not C^3).

Let us examine conformal infinity of the Schwarzschild solution. The familiar form of the metric is

$$ds^2 = (1 - \frac{2m}{r})\,dt^2 - \frac{dr^2}{1 - \frac{2m}{r}} - r^2(d\theta^2 + \sin^2\theta d\phi^2)\ .$$

Rather than attempt to obtain I^+ and I^- simultaneously, as was done for Minkowski space, it is simpler to introduce a retarded time coordinate

$$u = t - r - 2m \log (r - 2m)$$

and an advanced time coordinate

$$v = t + r + 2\,m \log (r - 2m)$$

separately. In the first case the metric becomes

$$ds^2 = (1 - \frac{2m}{r})du^2 + 2dudr - r^2(d\theta^2 + \sin^2\theta d\phi^2) \qquad (2.6)$$

and in the second

$$ds^2 = (1 - \frac{2m}{r})dv^2 - 2dvdr - r^2(d\theta^2 + \sin^2\theta d\phi^2)\ . \qquad (2.7)$$

In each case we can choose $\Omega = r^{-1} = w$, say.
The "unphysical metric" is

$$d\hat{s}^2 = \Omega^2 ds^2 = (w^2 - 2mw^3)du^2 - 2dudw - d\theta^2 - \sin^2\theta d\phi^2 \quad (2.8)$$

in the first case and

$$d\hat{s}^2 = (w^2 - 2mw^3)dv^2 + 2dvdw - d\theta^2 - \sin^2\theta d\phi^2 \qquad (2.9)$$

in the second. The metrics (2.8) and (2.9) are manifestly regular (and analytic) on their respective hypersurfaces $w = 0$. (Clearly the determinants are non-zero at $w = 0$.) The physical space-time is given when $w > 0$ in (2.8) and we can extend the manifold to include the boundary hypersurface I^+, given when $w = 0$. Similarly, in (2.9), the physical space-time has $w > 0$ and can be extended to include I^-, given when $w = 0$. In fact, we could if desired extend the space-time *across* $w = 0$ to negative values of w, but this will not be done here. Only the boundary $I = I^- \cup I^+$ will be adjoined to the space-time.

It is worth noting, at this point, a difficulty that is encountered if we try to identify I^- and I^+. If we do extend the region of definition of (2.8) to include negative values of w, then making the replacement $w \rightarrow -w$ we see that the metric has just the form (2.9) (with u in place of v) but where the mass m has been replaced by $-m$. Thus, the extension across I involves a reversal of the sign of the mass. In fact the derivative, at I, of the (conformal) curvature contains the information of the mass. (I shall elaborate on this point later, cf. (3.44).) It follows, therefore, that if we attempt to identify I^+ with I^-, with the *same* sign of the (non-zero) mass occurring on the two sides, then there must be a discontinuity in the derivative of the curvature across I (so that the metric $d\hat{s}$ must fail to be C^3 at I).

Accepting, then, that it is not reasonable to identify I^- with I^+, we are led to a picture closely resembling the one we obtained earlier in this section for Minkowski space. The only essential difference occurs with the points I^-, I°, I^+. It turns out that with mass present, the point I°, and normally $I^\pm$ also, must, if adjoined to the manifold, be *singular* for the conformal geometry. (I shall not go into the argument here.) It is therefore reasonable not to attempt to include these points, in the general case, as part of the conformal infinity. The picture, then, is as indicated in Figure 8. We have two disjoint boundary hypersurfaces I^- and I^+ each of which is a "cylinder" with topology $S^2 \times \mathbb{R}$. It is clear from (2.8) and (2.9) that each of $I^\pm$ is a *null* hypersurface (the induced metric, at $w = 0$, being degenerate). These null hypersurfaces are generated by null geodesics (given by θ, ϕ = const., $w = 0$) the tangents to which are the *normals* to the hypersurfaces. (Being null, the normals are also tangential.) These null geodesics may be taken to be the " $\mathbb{R}$'s" of the topological product $S^2 \times \mathbb{R}$ (i.e. the inverse images of the points of S^2 in the natural projection $S^2 \times \mathbb{R} \rightarrow S^2$).

We have obtained this structure explicitly in the case of the Schwarzschild metric. But it is clear that many other suitably asymptotically flat space-times will also give rise to such a structure. Let us start from a metric of the form

$$ds^2 = r^{-2}A\,dr^2 + 2B_i\,dx^i dr + r^2 C_{ij}\,dx^i dx^j \;, \qquad (2.10)$$

the coordinates being r, x^1, x^2, x^3. Each of A, B_i, C_{ij} is a suit-

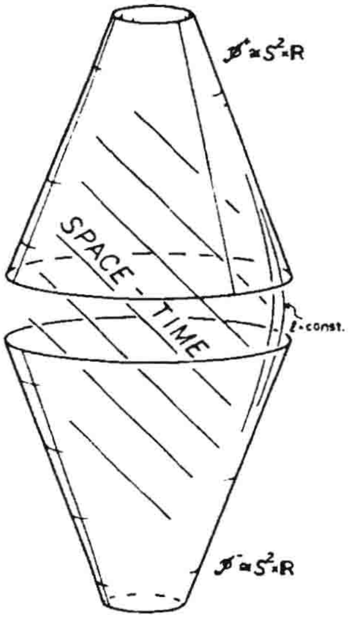

$\mathscr{I}^+ \cong S^2 \times R$
SPACE-TIME
$\ell = $ const.
$\mathscr{I}^- \cong S^2 \times R$

Fig. 8. Conformal infinity for an asymptotically flat space-time (e.g. the Schwarzschild solution). The points I^-, I°, I^+ have become singular and so are deleted; $\mathscr{I}^-$ and $\mathscr{I}^+$ are null cylinders $\cong S^2 \times \mathbb{R}$.

ably smooth function of x^0, x^1, x^2, x^3, with $x^0 = r^{-1}$ - these functions being smooth also at $x^0 = 0$. Then setting $\Omega = r^{-1}$ we have

$$d\hat{s}^2 = \Omega^2 ds^2 = A(dx^0)^2 - 2B_i dx^i dx^0 + C_{ij} dx^i dx^j \ . \qquad (2.11)$$

Provided the relevant determinant formed from A, B_i and C_{ij} does not vanish, the metric (2.11) will be perfectly regular at $x^0 = 0$. Thus, a "conformal infinity" will exist for the space-time whose metric is given by (2.10).

Many metrics used in the study of gravitational radiation do in fact have the form (2.10). In particular, this applies to the metrics of Bondi (and his coworkers) [11], and Sachs [12], Robinson-Trautman and Newman-Unti [13]. These metrics describe a situation where there is an isolated source (with asymptotic flatness) and outgoing gravitational radiation. Incoming gravitational radiation (of a suitably curtailed duration) may also be present. So may non-gravitational (i.e. electromagnetic or neutrino) zero rest-mass radiation. For such a situation, therefore, we expect a future-null conformal infinity $\mathscr{I}^+$ to exist. For the time-reversed situation we would expect $\mathscr{I}^-$ to exist. There will also be a wide class of "physically reasonable" situations in which both $\mathscr{I}^+$ and $\mathscr{I}^-$ exist. I feel there is likely to be no real loss of generality involved in the assumption that both $\mathscr{I}^+$ and $\mathscr{I}^-$ should exist for an asymptotically flat space-time. It has occasionally been argued [9] that the assumption of the existence of I^- may impose unnecessarily severe restrictions on the behaviour of the outgoing radiation in the infinite past. However, I do not think that this is really so. One of the major difficulties in gravitational radiation theory which one encounters, if an approach such as the one I am setting forth here is *not* adopted, is to find a reasonably clear-cut coordinate independent *definition* of what one should mean by incoming (or outgoing) gravitational radiation. If I^- exists, then the incoming field can be recognized in terms of the curvature at I^- (cf. later). If I^- does not exist, then the situation is much less clear. Also, it seems that both I^+ and I^- do in fact exist for "finite" scattering problems, for which a finite number of particles come in and go out along hyperbolic-type orbits. There seems no reason why I^- and I^+ should not exist in many other "reasonable" scattering problem situations - although examples can obviously be concocted involving infinite wave trains in which either or both of $I^{\pm}$ fail to exist. Whether such examples are regarded as "physically reasonable" is clearly a matter of taste. Asymptotic flatness is, in any case, a mathematical idealization. I feel this is a subject worthy of further study, however.

Asymptotically flat space-times of the type I am considering constitute the most important subclass of those space-times which are termed "asymptotically simple" (or, more generally, "weakly asymptotically simple"). A space-time M is *asymptotically simple* [6] if a conformal factor Ω exists for which the metric $d\hat{s} = \Omega ds$ remains smooth (say C^4) on some extension $\bar{M}$ of the manifold M which includes a boundary I, Ω being smooth (say C^4) throughout $\bar{M} = M \cup I$, becoming zero on I and having non-zero gradient at I, the global assumption being also made that every maximal null geodesic in M acquires both a past and a future end-point on I. It is sometimes convenient to weaken the global assumption that all maximal null geodesics should reach conformal infinity both in the future and in the past. This is in order to admit the possibility that some null geodesics might not escape to conformal infinity (as would be the case with a black hole or some other similar situations). The space-time M is said to be *weakly asymptotically simple* [8] if, roughly speaking, it possesses the conformal infinity of an asymptotically simple space-time, but it may possess other "infinities" as well. More precisely, the condition is that some asymptotically simple M' should exist such that for some neighbourhood K' of I' in $\bar{M}'$, the portion $K' \cap M'$ should be isometric with a subset of M.

The boundary I must in all cases be a smooth hypersurface. If it happens to be null, then it can be shown [6] to have the same topological structure that we have obtained for the Minkowski and Schwarzschild cases (cf. Figure 8). Furthermore, if Einstein's vacuum equations without cosmological term (for the physical metric ds) are assumed to hold *near* I (i.e. throughout $K-I$, where K is some neighbourhood of I in $\bar{M}$), then it follows that I *is* everywhere null. The same conclusion also follows under much weaker assumptions. For example, near I, we might impose Einstein's equations without cosmological term, with zero rest-mass fields as source. If, on the other hand, a cosmological term is present, then I is spacelike or timelike according as the cosmological constant is positive or negative.

These properties are simple consequences of the transformation formula for the Ricci tensor. Let us write this formula as

$$P_{ab} = \hat{P}_{ab} + \Omega^{-1}\hat{\nabla}_a\hat{\nabla}_b\Omega - \tfrac{1}{2}\,\Omega^{-2}\hat{g}_{ab}\hat{g}^{cd}\hat{\nabla}_c\Omega\hat{\nabla}_d\Omega \tag{2.12}$$

where

$$P_{ab} = \frac{1}{12}\,R\,g_{ab} - \frac{1}{2}\,R_{ab} \tag{2.13}$$

is a tensor (often occurring in conformal transformation expressions) containing the same information as the Ricci tensor R_{ab} for the ds metric, where $\hat{P}_{ab}$ is the corresponding tensor for $d\hat{s}$, and where $\hat{\nabla}_a$ denotes covariant derivative for the $d\hat{s}$ metric. Since $d\hat{s}$ and Ω are regular on I, the quantities $\hat{P}_{ab}$, $\hat{g}_{ab}$, $\hat{g}^{ab}$, $\hat{\nabla}_a\Omega$ and $\hat{\nabla}_a\hat{\nabla}_b\Omega$ all remain finite and continuous at I. Now if Einstein's vacuum equations hold without cosmological term (near I) we have $R_{ab} = 0$. This is equivalent to $P_{ab} = 0$. Thus, multiplying (2.12) by Ω^2 and using the condition $\Omega = 0$ on I, we obtain the fact that the vector

$$\hat{n}_a = \mp\hat{\nabla}_a\Omega \tag{2.14}$$

is null at $I^{\pm}$ (sign chosen so that $\hat{n}_a$ is future-pointing). But $\hat{n}_a$ is normal to Ω = const.; in particular, it is normal to I (and does not vanish at I, by one of the assumptions of asymptotic simplicity). Hence I is a null hypersurface. This same conclusion will also follow if we assume merely that the trace R of the physical Ricci tensor vanishes near I (the normal situation if all fields present are massless and there is no cosmological term). For we need only apply the above argument to the transvection of (2.12) with $\hat{g}^{ab}$ (i.e. to the "trace" of (2.12)) and the result is the same. Furthermore, if a cosmological constant is present, essentially the same argument will yield the other results mentioned above.

As a final consequence of (2.12), let us derive one further geometrical property of the hypersurface I. For simplicity we assume that there is vacuum near I, and no cosmological constant. The same property would actually follow under weaker assumptions (e.g. Einstein - Maxwell equations, without cosmological term, near I), but this is more difficult to derive. So take $P_{ab} = 0$ near I, and consider the "trace-free" part of (2.12) (i.e. transvect with $\delta^a_u\delta^b_v - \frac{1}{4}\,\hat{g}^{ab}\hat{g}_{uv}$). Multiplying by Ω and setting $\Omega = 0$, we obtain the equation

$$\hat{\nabla}_a\hat{n}_b = \{\tfrac{1}{4}\,\hat{g}^{cd}\hat{\nabla}_c\hat{n}_d\}\hat{g}_{ab} \quad \text{at} \quad I \tag{2.15}$$

(cf. (2.14)). Equation (2.15) tells us that the generators of I are non-rotating and shear-free. (These generators are the integral curves of normal vectors $\hat{n}^a$ to I - which, being null, are also tangent vectors to I.) The conclusion that they are non-rotating is trivial since I is a null hypersurface (symmetry of $\hat{\nabla}_a\hat{n}_b$). On the other hand, the shear-free nature of the generators ($\hat{\nabla}_a\hat{n}_b$ trace-free) tells us that small shapes are preserved as we follow these generators along I. That is to say, if we take any two cross-sections S_1, S_2 of I^+ or of I^-, then the correspondence between S_1 and S_2 established by the generators is a *conformal* one. This is the same situation as we encountered earlier, in Section 1, with the light cone of a point (cf. Figure 2). It implies that the space of generators of I^+ or of I^- has a conformal structure.

The fact that I^+ and I^- have this type of structure is essential for the definition of the B.M.S. group, as we shall see in the next section. We can take as the appropriate definition of asymptotic flatness, therefore, that a conformal infinity I (or at least a I^+) should exist with the structure as defined above. Weak asymptotic simplicity, together with Einstein-Maxwell equations (without cosmological term) holding near I will be sufficient to ensure this.

3. THE B.M.S. GROUP

In special relativity, the Poincaré Group arises in a natural way, as the group of symmetries of (Minkowski) space-time preserving its metric structure. In general relativity, however, no interesting group arises as the group of symmetries of general space-time. (I do not consider the "general coordinate group" or equivalently, the "group of diffeomorphisms of space-time" to be an interesting group in the context of general relativity theory. Since it preserves only *smoothness*, its relevance would be as much to any other space-time theory involving differential equations as to general relativity. And the group consisting of the identity element alone is even less interesting!) If we restrict attention to space-times which are asymptotically flat, however, then an interesting new concept does arise, namely that of an *asymptotic symmetry group*. In their original approach to the problem, Bondi and Metzner [11], and later Sachs [12,14], defined their asymptotic symmetry group (the B.M.S. group) as a group of transformations between asymptotically flat coordinate systems of a certain type. However, the geometrical approach to asymptotic flatness described in the last section affords us a much more vivid picture of the significance of this group as a transformation group and of its relevance in the presence of gravitational radiation [10].

Consider Minkowski space first. Since the transformations belonging to the Poincaré group are metric preserving, they are certainly conformal. They therefore induce transformations of I to itself which are also, in the appropriate sense, conformal (I being invariant). The Poincaré group may thus be thought of as a certain group of conformal symmetries of the Minkowski conformal infinity. In the same way, we might expect to be able to identify the Poincaré group as a group of conformal symmetries of the conformal infinity I of an asymptotically flat space-time; or we might expect that there would be some obstruction to making such an identification.

We must, however, be somewhat careful when it comes to defining exactly what structure on I is to be preserved by the transformations of this group. The most obvious structure that we can assign to I is its "inner conformal metric". But, I being a null hypersurface, this turns out to be rather too weak a structure. Distances along generators of I are just zero, so their ratios cannot be assigned. The group of self-transformations of I preserving this weak inner conformal metric does, however, have some significance as an asymptotic symmetry group. Like the B.M.S. group, it is the same group for every asymptotically flat space-time. It is considerably larger than the B.M.S. group and is referred to as the Newman-Unti group [13]. Before considering the strengthening of the structure of I that is required for the B.M.S. group, it is worthwhile to examine the nature of the inner conformal metric of I and these self-transformations which preserve it.

In Section 2 the structure of I for an asymptotically flat space-time M was described as the disjoint union of two smooth null hypersurfaces I^- and I^+ (where I^- consists of past end-points of null geodesics in M and I^+ of future end-points), the topology of each of $I^\pm$ being $S^2 \times \mathbb{R}$, where the "$\mathbb{R}$'s" may be taken as the null-geodesic generators of $I^\pm$. These generators are shear-free and so establish a conformal mapping between any two S^2 cross-sections of I^+ or of I^-. Now it is a theorem that any conformal 2-surface with the topology of a sphere S^2 is in fact conformal to the unit 2-sphere in Euclidean 3-space. Thus, without loss of generality we can assume that the conformal factor Ω has been chosen so that some cross-section S of I^+, say, has the (unphysical) metric $-d\hat{s}^2$ of a unit 2-sphere. For, given one choice of Ω, we can always make a new choice $\Omega' = \Theta\,\Omega$ which again has the required properties of vanishing at I with non-zero gradient there, the factor Θ being an arbitrary smooth positive function on I which may be taken to rescale the metric on I as we please. We can in fact use this freedom in Ω to scale the metric $d\hat{s}$ along every generator of I^+ so that the divergence of these generators vanishes. Put another way, we arrange that a continuous succession of cross-sections of I^+ have metrics agreeing with that of S, as mapped along the generators. (These cross-sections are already conformal to S by this map.) Thus, without loss of generality, we can assume the (unphysical) metric of I^+ to be given by

$$d\ell^2 = -\,d\hat{s}^2 = d\theta^2 + \sin^2\theta d\phi^2 + 0.du^2 \tag{3.1}$$

where θ and ϕ are spherical polar coordinates for S, chosen to be constant along the generators of $\mathcal{I}^+$, and where u is a parameter defined along each generator of $\mathcal{I}^+$ (increasing monotonically with time from $-\infty$ to ∞) the surfaces u = const. being cross-sections of $\mathcal{I}^+$. It is clear from (3.1) that every cross-section (given by u = a function of θ, ϕ) of $\mathcal{I}^+$ now has the metric of a unit 2-sphere.

We can also choose Ω so that $\mathcal{I}^-$ has a metric of the form (3.1) as well, but where we use a coordinate v in place of u, with v also increasing into the future. Then u is a retarded time coordinate on $\mathcal{I}^+$ and v an advanced time coordinate on $\mathcal{I}^-$. These coordinates, and θ and ϕ also can, if desired, be extended into the finite space-time M. But this serves no real purpose here - from the point of view solely of asymptotic symmetry groups.

Let us consider the group of self-transformation of $\mathcal{I}^+$ which preserves its inner conformal metric. (I shall ignore time-reversing transformations. These should, strictly speaking, interchange $\mathcal{I}^+$ with $\mathcal{I}^-$.) Note first that any smooth transformation of $\mathcal{I}^+$ to itself which sends each generator into itself (and preserves the orientation on each generator) will be allowable. In terms of θ, ϕ, u these are given by

$$\begin{aligned} \theta &\to \theta \\ \phi &\to \phi \\ u &\to F(u, \theta, \phi) \end{aligned} \tag{3.2}$$

where F is smooth with $\partial F/\partial u > 0$. Clearly (3.1) is unchanged. In addition, we can allow conformal transformations of the (θ, ϕ)-sphere to itself. These were discussed in Section 1. For convenience we introduce

$$\zeta = e^{i\phi}\cot\frac{\theta}{2}\ , \tag{3.3}$$

as in (1.10), and the metric (3.1) becomes

$$d\ell^2 = \frac{4d\zeta d\bar{\zeta}}{(1+\zeta\bar{\zeta})^2} + 0.du^2 \tag{3.4}$$

(cf. (1.11)). The conformal maps of the sphere are now given by (1.12)(or this composed with $\zeta \to \bar{\zeta}$), so the general Newman-Unti transformation takes the form

$$\begin{aligned} \zeta &\to \frac{\alpha\zeta+\beta}{\gamma\zeta+\delta} \\ u &\to F(u, \theta, \phi) \end{aligned} \tag{3.5}$$

($\partial F/\partial u > 0$) or, in the case of spatially-reflective transformations, (3.5) composed with $\zeta \to \bar{\zeta}$.

There is a great deal of freedom in the function F and we want to be able to reduce this. In fact it is possible to assign rather more geometric structure to $\mathcal{I}^+$ than we have given hitherto. The preservation of this additional structure will reduce the freedom in F to that of a function of θ and ϕ only. This will give us the B.M.S. group. To obtain this additional structure, let us consider the geometrical interpretation of a point of $\mathcal{I}^+$. Assume, first, that M is Minkowski space. We saw that in this case each point p of $\mathcal{I}^+$ is associated with a null hyperplane π in M. Every generator of π is a null geodesic and these all attain the same future end-point p on $\mathcal{I}^+$. In physical terms, we may think of π as a constant phase hypersurface in a plane wave. The other constant phase hypersurfaces belonging to the same plane wave will be null hyperplanes parallel to π. These other null hyperplanes will terminate at other points on $\mathcal{I}^+$, but the totality of all these points, for the given plane wave (i.e. for a given null direction in Minkowski space) will constitute the generator γ of $\mathcal{I}^+$ through p. (Parallel null geodesics terminate on $\mathcal{I}^+$ at points of one generator of $\mathcal{I}^+$.) The different null geodesics through p (apart from γ) are the different generators of π. Now for fixed π the space of these generators has the structure of a Euclidean plane E^2. This is because any cross-section (in particular, any cross-section by a space-like plane) of π has the intrinsic metric of a Euclidean plane, and the projection along the generators of π maps these planes isometrically. (As an exercise verify these facts.) In the neighbourhood of p, these null geodesics generate the past light cone of p. We can see the Euclidean E^2 structure arising if we take a "parabolic" section of the past null cone of p by a null 3-plane near p which is parallel to γ (see Figure 9). The situation is similar to that depicted in Figure 3 (except that here we are choosing a past rather than a future null cone). The $-d\hat{s}^2$ metric on this section is that of a Euclidean plane - conformal to E^2. This correspondence will be made more explicit shortly.

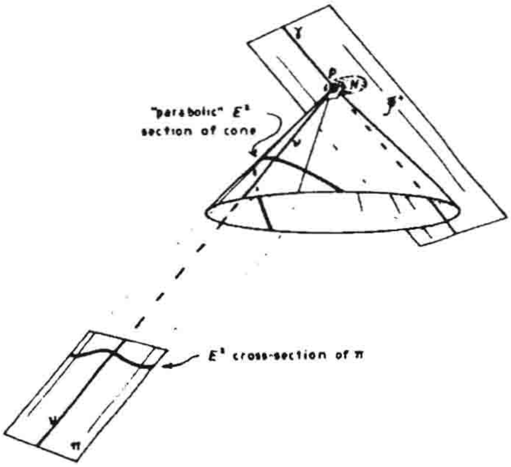


Fig. 9. The null hyperplane π becomes the past light cone of p. The generator ν is represented by an orthogonal 2-plane N tangent to $\mathcal{I}^+$ at p.

Now, by taking orthogonal compliments, we can refer this E^2 structure of the generator space of π to the *tangent space to $\mathscr{I}^+$ at* p. We associate any generator ν of π (i.e. null geodesic through p) with the 2-plane element at p which is spanned by the directions of γ and of ν at p. Since γ is normal to $\mathscr{I}^+$ at p, the orthogonal compliment of this 2-plane element will be another 2-plane element N, which is now *tangent* to $\mathscr{I}^+$ (and which does not contain the direction of γ). The 2-plane element N at p is characterized uniquely by the fact that it is tangent to $\mathscr{I}^+$ and orthogonal to ν. Thus, we may use tangent 2-plane elements to $\mathscr{I}^+$, to represent null geodesics in *M*. (This will also work if *M* is curved and asymptotically flat.) If we consider a 3-plane element, at p, which contains the direction of γ, we see that the intersection with the null cone at p gives us a one-dimensional system of null geodesics in π. It is not hard to see, *M* being Minkowski space, that this system corresponds to a *straight line* in E^2. Furthermore, the orthogonal compliment of this 3-plane element at p is a *line element* at p which is tangent to $\mathscr{I}^+$ (and not in the direction of γ). Thus line elements at p represent straight lines in E^2. *Directions* tangent to $\mathscr{I}^+$ at p represented *oriented* straight lines in E^2.

With very little change, this discussion can in fact be carried over to an asymptotically flat *M*. The past light cone of a point p on $\mathscr{I}^+$ (i.e. the locus of null geodesics in *M* which terminate at p) is a null hypersurface π in *M* which is *asymptotically* plane in the future. (Physically, this is a constant phase hypersurface of an outgoing asymptotically plane wave. The different constant phase hypersurfaces for one wave will be the past light cones with vertices on one generator of $\mathscr{I}^+$.) The cross-sections of π will not now be Euclidean planes exactly. Nor will the different cross-sections be isometric with one another. On the other hand, by taking the *limit* of such cross-sections as they recede into the future, we can recover an exact Euclidean plane E^2 to represent the space of generators of π. This may be seen by considering Figure 9 again. Our "parabolic" section of the past cone of p may be taken as close to $\mathscr{I}^+$ as we please - indeed we may consider it to be actually in the tangent space at p. Then we get an exact E^2 structure. This parabolic section can be given, in the *tangent space*, by the equation

$$\hat{x}^a \hat{n}_a = -1 \qquad (3.6)$$

$\hat{x}^a$ being the position vector of a point on the section (and subject also to $\hat{x}^a\hat{x}^b\hat{g}_{ab} = 0$, so that it lies on the null cone), the null vector $\hat{n}_a$ being given, as in (2.14), by

$$\hat{n}_a = -\hat{\nabla}_a \Omega \qquad (3.7)$$

at p. Now consider a point p' on π which lies in the remote future along π. In terms of *M*, p' will be a point lying just to the past of p on the light cone of p. We can (using "classical" notation) label p' by the position vector $d\hat{\ell}^a$, relative to p. The conformal factor at p' is thus

$$d\Omega = d\ell^a \hat{\nabla}_a \Omega \ , \qquad (3.8)$$

since $\Omega = 0$ at p. To pass from "unphysical" to "physical" distances at p' we must therefore divide by the factor (3.8). This is equivalent to expanding by the factor $\hat{\ell}^a : d\hat{\ell}^a$, where $\hat{\ell}^a$ is chosen in the same null direction from p as p', but subject to (3.6) (with (3.7)). It follows that to measure distances between points on π which lie in the remote future (i.e. "near" to p) we can simply refer to the "parabolic" section E^2, given by (3.6), it being ultimately relevant only on which generator of π the point lies, not how "far" it is up this generator. Hence, this E^2 does describe the geometry of cross-sections of π, but taken in the limit as these cross-sections recede into the remote future.

As remarked above, *null geodesics* in M can be represented by *2-plane elements tangent to* $\mathscr{I}^+$, similarly to the case for Minkowski space. Furthermore, a *line element* tangent to $\mathscr{I}^+$ at p (which is not in the direction of γ) represents a one-dimensional system of generators of π - of the special type which corresponds to a *straight line in* E^2 , so asymptotically they generate a null 2-plane in π. A tangent *direction* to $\mathscr{I}^+$ at p corresponds to an *oriented* straight line in E^2.

Let us now consider the relevant structure of $\mathscr{I}^+$. The conformal geometry of $\mathscr{I}^+$ affords us a definition of the *angle* between two line elements tangent to $\mathscr{I}^+$ at p. Representing each line element by a straight line in E^2 in accordance with the above, we can interpret this angle simply as the Euclidean angle between the two corresponding straight lines in E^2. Provided the two line elements at p do not span a 2-plane containing the direction of γ, this will be an ordinary finite Euclidean angle between two intersecting straight lines in E^2. In special cases, however, the two straight lines will become parallel. Then the angle between the corresponding line elements becomes zero. This is the situation occurring when the 2-plane spanned by the line elements *does* contain the direction of γ (neither of the line elements being actually *in* the direction of γ). Although the *angle* is zero, there is still an invariant concept of separation between such line elements. I shall call this concept a *null angle* [10]. A null angle has units of a *distance* (or equivalently a *time*) since it describes the distance in E^2 between the associated parallel straight lines (Figure 10). Thus, we may attribute to $\mathscr{I}^+$ a physically meaningful *strong conformal geometry*. In addition to the concept of *angle* between (non-null) tangent directions at a point of $\mathscr{I}^+$ (which concept is the content of the inner conformal metric of $\mathscr{I}^+$) the strong conformal geometry assigns a measure of separation whenever this angle becomes zero, termed the null angle between these directions, and having the dimensions of a distance (or time).

There is another (equivalent) way of specifying the strong

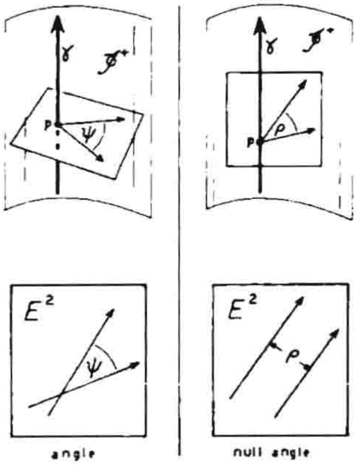


Fig. 10. The inner conformal metric of $\mathcal{I}^+$ assigns a measure of angle between any two (non-null) tangent directions to $\mathcal{I}^+$. When this angle is zero, the strong conformal geometry of $\mathcal{I}^+$ is needed to define the null angle between the directions. Representing directions on $\mathcal{I}^+$ by straight lines in E^2 we obtain finite angles as the angles between the lines, and null angles as the distance between them when these angles reduce to zero.

conformal geometry of $\mathcal{I}^+$. Observe that if we replace Ω by $\Omega' = \Theta\Omega$ (Θ being smooth and positive on $\bar{M}$), the normal vector $\hat{n}^a = \hat{g}^{ab}\hat{\nabla}_b\Omega$ is replaced by

$$\hat{n}'^a = \hat{g}'^{ab}\hat{\nabla}'_b\Omega' = \Theta^{-2}\hat{g}^{ab}\hat{\nabla}_b(\Theta\Omega) = \Theta^{-1}\hat{g}^{ab}\hat{\nabla}_b\Omega = \Theta^{-1}\hat{n}^a \qquad (3.9)$$

on $\mathcal{I}^+$ (since $\Omega = 0$). Hence (setting $d\ell^2 = -d\hat{s}^2$ for the metric on $\mathcal{I}^+$),

$$\hat{n}^a d\ell = \hat{n}'^a d\ell' \qquad (3.10)$$

so the quantity $\hat{n}^a d\ell$ (or $\hat{n}^a\hat{n}^b d\hat{s}^2$ if preferred) defines an invariant structure on $\mathcal{I}^+$. The vectors $\hat{n}^a$ are tangent to the generators of $\mathcal{I}^+$ and define a scaling for special parameters u along these generators according to

$$\frac{\partial}{\partial u} = \hat{n}^a\hat{\nabla}_a \ , \quad \text{i.e.} \quad \hat{n}^a\hat{\nabla}_a u = 1 \qquad (3.11)$$

(the operators acting on scalars). The invariance of $\hat{n}^a d\ell$ may be restated as the invariance of the ratio $du:d\ell$. This ratio is, in effect, a null angle (see Figure 11), the invariance of the ratios of the $d\ell$'s in different directions giving the inner conformal metric also.

If we make some specific choice of metric $d\ell$ for cross-sections of $\mathcal{I}^+$, then this singles out a specific scaling for the parameter u. It is usual to choose the metric $d\ell$ to be that of a unit

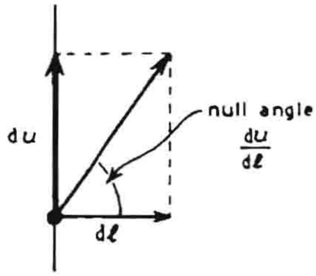


Fig. 11. The ratio $du:d\ell$ defines a null angle.

sphere, so that the entire metric for $\mathcal{I}^+$ takes the form (3.1). The scalings for the associated parameters u are now fixed. The value of the null angle $du/d\ell$ (in seconds, say) defines the scaling for u (in seconds). The only arbitrariness in u now lies in fixing the origin of the u-coordinate on each generator of $\mathcal{I}^+$.

A transformation of $\mathcal{I}^+$ to itself which preserves $\hat{n}^a d\ell$ - or, equivalently, which preserves angles and null angles - must have the effect that any expansion (or contraction) of the spatial distances $d\ell$ is accompanied by an equal expansion (or contraction) of the scaling of the special u-parameters. The allowed transformations must have the form (3.5) (since these are the ones preserving the inner conformal metric), but the function F must now have the special form that allows the ratio $du:d\ell$ to remain invariant. The sphere of cross-section of $\mathcal{I}^+$ undergoes a conformal mapping, so

$$d\ell \to K d\ell$$

where K is a positive function of the angular coordinates θ, ϕ or ζ, $\bar{\zeta}$. We must therefore also have

$$du \to K\, du.$$

Since K is independent of u, this means u transforms according to

$$u \to K(u + a(\zeta,\bar{\zeta})) \tag{3.12}$$

where a is some real function defined on the ζ-sphere. Using (3.4) and the explicit form (3.5) for the conformal transformations of the sphere, we obtain

$$K = \frac{1 + \zeta\bar{\zeta}}{(|\alpha\zeta+\beta|^2 + |\gamma\zeta+\delta|^2)}$$

α, β, γ, δ being complex parameters subject to

$$\alpha\delta - \gamma\beta = 1 \tag{3.13}$$

The general form of the transformation is then

$$\zeta \rightarrow \frac{\alpha\zeta+\beta}{\gamma\zeta+\delta}$$
$$u \rightarrow \frac{(1+\zeta\bar{\zeta})\{u+a(\zeta,\bar{\zeta})\}}{(|\alpha\zeta+\beta|^2+|\gamma\zeta+\delta|^2)} \tag{3.14}$$

where the arbitrary function $a(\zeta,\bar{\zeta})$ may be taken to be suitably smooth on the sphere*. The transformations (3.14) (or these composed with $\zeta \rightarrow \bar{\zeta}$) define the *Bondi-Metzner-Sachs* (or B.M.S.) group, this being the *group of self-transformations of* $\mathcal{I}^+$ *which preserves strong conformal geometry*.

The particular transformations of the form

$$\zeta \rightarrow \zeta$$
$$u \rightarrow u+a(\zeta,\bar{\zeta}) \tag{3.15}$$

are called *supertranslations*. Here each generator of $\mathcal{I}^+$ is sent into itself, being simply shunted along itself by an amount $a(\zeta,\bar{\zeta})$. Among these transformations are the ones called *translations* for which a is composed solely of spherical harmonics of orders 0 and 1. In terms of $\zeta,\bar{\zeta}$, these are

$$a(\zeta,\bar{\zeta}) = \frac{A+B\zeta+\bar{B}\bar{\zeta}+C\zeta\bar{\zeta}}{1+\zeta\bar{\zeta}} \tag{3.16}$$

where A, C are real and B complex. These transformations are in fact the ones which translations in Minkowski space induce on $\mathcal{I}^+$ (exercise). The supertranslations form an infinite parameter subgroup of the B.M.S. group and the translations form a four-parameter subgroup. In each case it is a *normal* subgroup. In fact, Sachs has shown [14] that the translation subgroup is the *only* four-parameter normal subgroup of the B.M.S. group; the supertranslation group is the largest proper normal subgroup of the B.M.S. group. Thus, each of these subgroups is singled out by its group-theoretic properties alone.

The factor group of the B.M.S. group by the supertranslation group is the conformal group on the sphere S^2. This sphere may be taken to be the space of generators of $\mathcal{I}^+$. As we saw in Section 1, this factor group is the orthochronous Lorentz group. Note that whereas the translation group occurs canonically as a *subgroup* of the B.M.S. group, we are obtaining the Lorentz group canonically only as a *factor* group. We cannot fit these two concepts together to obtain the Poincaré group in a canonical way from the B.M.S.

---

* There is a certain amount of choice as to the degree of smoothness or continuity taken for $a(\zeta,\bar{\zeta})$. This can make a considerable difference when it comes to the representation theory of the B.M.S. group. Representation theory for the B.M.S. group has been studied by Sachs [14], Cantoni [15,16] and McCarthy [17,18,19, 20].

group.

The Lorentz group occurs here as a factor group of the B.M.S. group by the infinite-parameter Abelian group of supertranslations. In the case of the Poincaré group, the Lorentz group appears as a factor group of it by the four-parameter Abelian group of translations. Each of the two groups, the B.M.S. group and the (orthochronous) Poincaré group, is what is known as a semidirect product [17,21] of the (orthochronous) Lorentz group with the appropriate (super) translation group. However, the fact that the translation group sits in the supertranslation group does not, in itself, allow us to find the (orthochronous) Poincaré group as a subgroup of the B.M.S. group in a canonical way. The difficulty occurs with defining what would be meant by a "supertranslation-free" Lorentz transformation.

Let us revert to Minkowski space M and see how the Poincare group arises in that case as a subgroup of the B.M.S. group. It should be clear that the orthochronous Poincaré group on this M does in fact arise as a subgroup of the B.M.S. group on I^+. For the strong conformal geometry of I^+ was defined entirely in terms of the metric structure of the space-time. Thus, any self-transformation of M preserving this metric structure (and time-sense) must also preserve the strong conformal geometry of I^+. However, the B.M.S. group, being much larger than the Poincaré group, preserves less structure on I^+ than does the Poincaré group. Let us examine the nature of this additional structure of I^+ whose preservation, in the case of Minkowski space M, allows us to restrict the B.M.S. transformations to Poincaré transformations.

The significance of the Poincaré transformations in this context is, of course, that they can be applied actually to the points of M, not just to I^+. So the question arises as to how to recognize points of M in terms of I^+. Clearly, the future light cone Q of any point $q \in M$ can be used to represent the point q. The intersection $\bar{Q} \cap I^+$ is a cross-section of I^+ which is itself sufficient to characterize Q and therefore the point q. (For $\bar{Q} \cap I^+$ is a spacelike 2-surface. Any spacelike 2-surface is locally the intersection of two uniquely defined null hypersurfaces. These null hypersurfaces are generated by the two systems of null geodesics which meet the 2-surface orthogonally. In the present situation these two null hypersurfaces are simply I^+ and $\bar{Q}$.) Let us call a cross-section of I^+ which arises in this way a good cross-section. A bad cross-section of I^+ is, on the other hand, the intersection of I^+ with some null hypersurface which does not come together cleanly at a single vertex [22]. Instead, the null geodesic generators of this hypersurface will encounter caustics and crossover regions somewhere in the interior of M (Figure 12). The B.M.S. transformations of I^+ which send good cross-sections into good cross-sections can be regarded as acting on the points of M. In fact they will be (orthochronous) Poincaré transformations, as we shall see.

We can use standard coordinates (r, u, θ, ϕ) for Minkowski

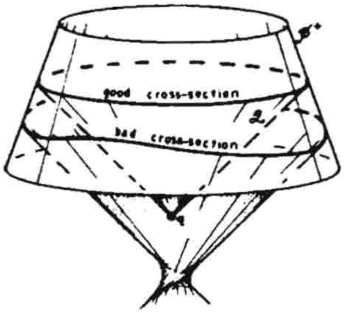


Fig. 12. A good cross-section of $\mathcal{I}^+$ for Minkowski space M is one arising as the intersection of $\mathcal{I}^+$ with the light cone of a point in M. The B.M.S. transformations sending good cross-sections to good cross-sections constitute the (orthochronous) Poincaré group on M.

space, as in (2.3), (2.4), (2.6), where r is radial, u a retarded time, and θ, ϕ spherical polar coordinates, the physical metric taking the form

$$ds^2 = du^2 + 2dudr - r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{3.17}$$

The parameter u is of the special type required for (3.11) as is readily verified. In fact the metric (3.17) is an example of the Bondi-Sachs form, which is (2.10) with $x^1 = u$, $x^2 = \theta$. $x^3 = \phi$ and

$$(C_{ij}) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -\sin^2\theta \end{pmatrix} + O(r^{-1}) \tag{3.18}$$

so that the $d\ell^2$ metric of $\mathcal{I}^+$ takes the form (3.1) (with $\Omega = r^{-1}$); and where

$$B_1 = 1 + O(r^{-1}), \tag{3.19}$$

from which it follows directly that u is of the special type required for (3.11). (For, by (2.14), we have $\hat{n}_a = -\delta_a^0$; also $\hat{g}^{0b} = -\delta_1^b$ by (3.18), (3.19), (2.11), so $\hat{n}^a = \delta_1^a$ giving (3.11).) A u-parameter of this kind, for which u-const. are null hypersurfaces in M, is sometimes referred to as a Bondi-type retarded time. It was by consideration of the coordinate group preserving (3.18) and (3.19) that Bondi and Metzner [11] (in the axi-symmetric case) and Sachs [12] (in the general case) were first led to the B.M.S. group.

Returning to the special case of Minkowski space and the special form (3.17), we see that each cross-section of $\mathcal{I}^+$ given by u = const. is a good cross-section, since it arises from the light cone of a point on the origin-axis r = 0. The remaining good cross-sections are obtained from u = 0 by means of the translations (3.15), (3.16), i.e.

$$u = \frac{A+B\zeta+\overline{B\zeta}+C\zeta\bar{\zeta}}{1+\zeta\bar{\zeta}} \tag{3.19a}$$

Setting $\zeta = e^{i\phi}\cot\frac{\theta}{2}$, we can re-express (3.19a) in terms of spherical polar coordinates

$$u = \left(\frac{A+C}{2}\right) + \left(\frac{C-A}{2}\right)\cos\theta + \left(\frac{B+\bar{B}}{2}\right)\sin\theta\cos\phi + \\ + i\left(\frac{B-\bar{B}}{2}\right)\sin\theta\sin\phi\ . \tag{3.20}$$

If we represent $\mathscr{I}^+$ as the spherical cylinder in $\mathbb{R}^4$ given by (u, sinθcosϕ, sinθsinϕ, cosθ) then the good cross-sections (3.20) are seen to be simply the sections of this cylinder by 3-planes in $\mathbb{R}^4$ (not parallel to the axis).

A (restricted) Lorentz transformation of *M*, about the origin r = 0, u = 0, leaves the particular good cross-section of $\mathscr{I}^+$ given by u = 0 invariant. Such a transformation - in fact the general (restricted) B.M.S. transformation leaving u = 0 invariant - is given by

$$\zeta \to \frac{\alpha\zeta+\beta}{\gamma\zeta+\delta}\ , \qquad u \to \frac{(1+\zeta\bar{\zeta})u}{|\alpha\zeta+\beta|^2+|\gamma\zeta+\delta|^2} \tag{3.21}$$

where αδ-βγ = 1. These transformations must send good cross-sections into good cross-sections, so they must preserve the form (3.19a) (as, indeed, is easy to verify). Now, the general (restricted) B.M.S. transformation which sends good cross-sections to good cross-sections must obtain the particular good cross-section u = 0 from some other good cross-section. We can therefore express the B.M.S. transformation as the composition of the translation which sends this other good cross-section into u = 0, with a Lorentz transformation (i.e. (3.21)) which leaves u = 0 invariant. Thus the B.M.S. transformation can be expressed as:

$$\zeta \to \frac{\alpha\zeta+\beta}{\gamma\zeta+\delta}, \qquad u \to \left(\frac{1+\zeta\bar{\zeta}}{|\alpha\zeta+\beta|^2+|\gamma\zeta+\delta|^2}\right)\left(u + \frac{A+B\zeta+\overline{B\zeta}+C}{1+\zeta\bar{\zeta}}\right) \tag{3.22}$$

the parameters α,β,γ,δ,B being complex, with αδ-βγ = 1, and A,C being real. The B.M.S. transformations sending good cross-sections into good cross-sections thus form a ten-real-parameter (connected) group. Since any restricted Poincaré transformation on *M* belongs to this group, the group (3.22) must actually be the restricted

Poincaré group on M.

There are, however, many other subgroups of the B.M.S. group which can be expressed in the form (3.22) and which are therefore isomorphic with the (restricted) Poincaré group. This arises from the fact that the group (3.22) is not a normal subgroup of the B.M.S. group; the transformations (3.22) do not commute with general supertranslations, so conjugate Poincaré subgroups with respect to supertranslations may be formed. Put another way, we may choose any other allowable u-coordinate, say u', where u' = 0 is a *bad* cross-section of I^+. (Thus, u' is obtained from u by a supertranslation which is not a translation.) Using u' in place of u in (3.22) we get another subgroup isomorphic with the restricted Poincaré group. To see that this subgroup is indeed different from the previous one, we may examine the "rotation group" subgroup for which each cross-section u' = const. is invariant (i.e., for which $A = B = C = \alpha\bar{\beta}+\gamma\bar{\delta} = 0$, $\alpha\bar{\alpha} = \delta\bar{\delta} = 1-\beta\bar{\beta} = 1-\gamma\bar{\gamma}$). The only (continuous) subgroups of the Poincaré group which are isomorphic with the rotation group are, in fact, the groups of rotation about some timelike line of origins in Minkowski space. Thus, if the above "rotation group" were to be contained in the Poincaré group on M, the invariant cross-sections of I^+ would all have to be good cross-sections (corresponding to the points of the timelike line of origins). Since all the u' = const. cross-sections are bad cross-sections and invariant, it follows that the copy of the Poincaré group that we have obtained is not the same as the Poincaré group on M.

In the general case the two groups will have only their translation subgroups in common. In particular cases, however, the bad cross-section u' = 0 may have axial symmetry. Then there will be rotations about the symmetry axis which are also common to the two groups. These particular rotations happen to be "supertranslation-free" in the sense that they send good cross-sections into good cross-sections, but even then most rotations will not have this property.

Let us turn, now, to the case when M is a general asymptotically flat space-time. The problem of identifying a particular subgroup of the B.M.S. group as *the* (restricted) Poincaré group may be phrased in terms of deciding which elements of the B.M.S. group are to be regarded as "supertranslation-free" rotations (or Lorentz rotations). As we have seen, the translations may be distinguished from the remaining supertranslations in a B.M.S.-invariant way; it is the "supertranslation-free" rotations which cannot be so distinguished from the general B.M.S. rotations. Since "supertranslation-free" rotations are to be ones sending "good" cross-sections into one another, it would seem that a suitable definition of "goodness" of cross-sections of I^+ is required.

The most obvious generalization, for an asymptotically flat space-time M, of the Minkowskian definition of a "good" cross-section, namely as the intersection of I^+ with the light cone of a point in M, is totally inappropriate. In the first place, there are many perfectly reasonable asymptotically flat space-times in which no cross-sections of $\mathcal{I}^+$ whatever would arise in this way. A model describing two stars a large distance from each other would be an example. If the stars are far enough apart, then every point q of the space-time would have the property that it is "beyond the focal length" of one or other of the stars. That is to say, the null geodesics from q would be focussed by that star to the extent that the light cone of q would encounter caustics and crossover regions behind the star. This would have the effect that when these null geodesics reach $\mathcal{I}^+$, they would not give a proper cross-section of $\mathcal{I}^+$, but a certain surface which wraps backwards and forwards on $\mathcal{I}^+$ intersecting some generators more than once.

However, the situation is considerably worse than this. Even if we restrict attention only to asymptotically flat space-times which do contain a reasonable number of "good" cross-sections of this kind, we are not likely to obtain any group of B.M.S. transformations (apart from the identity) which sends this system of cross-sections into itself, let alone a group isomorphic with the Poincaré group. The difficulty is that the detailed irregularities of the structure of the interior of the space-time would be reflected in the definition of "goodness" of cross-sections of $\mathcal{I}'$. It would seem that this whole approach is quite wrong. When discussing an asymptotic symmetry group, we should use asymptotic properties only (i.e. properties defined in the neightbourhood of $\mathcal{I}^+$ alone) in order to specify the structure we require.

There is a more satisfactory way, in the case of Minkowski space M, of recognizing whether or not a cross-section of $\mathcal{I}^+$ is "good". This is to examine the asymptotic shear of the null hypersurface $\bar{Q}$ which intersects $\mathcal{I}^+$ in the given cross-section. The concept of "shear", for null geodesics, was mentioned briefly at the end of Section 2. I shall be a little more explicit shortly, but for the moment let me just say that in any asymptotically flat space-time M containing a congruence (i.e. a three-dimensional system) Γ of null geodesics there is a complex quantity σ°, defined at the intersection with $\mathcal{I}^+$ of each null geodesic of Γ, whose vanishing is necessary and sufficient for the shear of Γ to vanish asymptotically (or, strictly speaking, to fall off faster than the inverse square of an affine parameter) [24], so that small shapes are preserved asymptotically as we follow the lines of Γ. The quantity σ° can also be defined for a single null hypersurface; $\sigma^\circ = 0$ states that the null hypersurface is asymptotically shear-free. Since the null hypersurface is determined by its intersection with $\mathcal{I}^+$, we can refer the definition of σ° simply to this cross-section on $\mathcal{I}^+$. Now, it can be shown that, in the case of Minkowski space, those cross-sections of $\mathcal{I}^+$ which are shear-free (i.e. $\sigma^\circ = 0$) are precisely the good cross-sections that we defined earlier. Thus a definition of "goodness" is provided, for Minkowski space, which refers only to quantities defined asymptotically.

However, this definition encounters difficulties also, when applied to asymptotically flat space-times. In the first place, suppose we have a cross-section S of $\mathcal{I}^+$ which is shear-free in the above sense. If this is to be considered as a "good" cross-section then every cross-section obtainable from S by means of a translation ought also to be considered as a "good" cross-section (since the translation subgroup ought to belong to the "Poincaré group"). But when gravitational radiation is present, most translations of S will *not* be shear-free. It turns out, in fact [11,12], that if we define σ° suitably for each u = const. cross-section of $\mathcal{I}^+$ (u being a Bondi-type retarded time) then $\partial^2\sigma^\circ/\partial u^2$ is a measure of the outgoing gravitational radiation field expressed in terms of the asymptotic curvature. The quantity $\partial\sigma^\circ/\partial u$ is the Bondi-Sachs "news function" whose squared modulus represents the flux of energy-momentum of gravitational radiation [10,11,12,13] across $\mathcal{I}^+$ (i.e. the energy-momentum which flows out of the system in the form of gravitational radiation). Thus, if S is given by $u = 0$ and is shear-free, the cross-sections u = const., which are translations of $\mathcal{I}^+$, will *not* be shear-free in the presence of gravitational radiation. I shall consider these formulae in a little more detail shortly.

There is a more serious difficulty even than this, however, namely that for a general asymptotically flat space-time, $\mathcal{I}^+$ will contain *no* shear-free cross-sections at all. This is to be expected because the freedom in choosing a cross-section of $\mathcal{I}^+$ is *one* real number per point of the section (i.e. the u-value on each generator of $\mathcal{I}^+$) whereas the quantity σ° is *complex*, its vanishing, therefore, representing *two* real numbers per point of the section. Only for particular space-times can shear-free cross-sections be found. One class of space-times for which such sections do exist is that consisting of all *stationary* asymptotically flat space-times. It is of interest that in this case it is *also* true that every translation of a shear-free cross-section is again shear-free, from which it follows that a particular Poincaré subgroup of the B.M.S. group *can* be singled out canonically as the group of B.M.S. transformations which send shear-free cross-sections of $\mathcal{I}^+$ into shear-free cross-sections. But in the general case, when the space-time is not stationary, this will not work. It may be remarked, for example, that in a Bondi axi-symmetric reflection-symmetric radiating space-time, a 2-parameter system of good cross-sections exists but it is not invariant under translations.

We can attempt to weaken the condition $\sigma^\circ = 0$ to something like "the real part of σ° vanishes on the cross-section". For technical reasons (σ° is a spin-weighted quantity [22]), we would have to put this a little differently and specify instead that, say, what is known as the "electric part" of σ° should vanish [22]. Such cross-sections do then exist, but the other difficulty remains. Such sections are not sent into one another by translations.

In this connection it is worth mentioning an interesting new approach to the whole question of asymptotic symmetry groups for asymptotically flat space-times, which has been inititated by Newman, and his collaborators [23]. In this approach, instead of asking for shear-free cross-sections of I^+ and for transformations between them, we ask instead for congruences of null geodesics for which $\sigma^\circ = 0$ and for transformations between them. In general, the rotation of such congruences will not vanish, so we do not get cross-sections of I^+ defined, or retarded time hypersurfaces. (We may, however, view the question as that of finding "complex cross-sections" of I^+. By allowing u to be complex, on a cross-section, we have the freedom, in effect, to make σ° zero). It seems possible, in this approach, that the Poincaré group (or a complex version of it) can be extracted as a kind of asymptotic symmetry group. The approach is, in a sense, complimentary to that of Bondi, Metzner and Sachs. It is closely related to the theory of twistors that will be described briefly in Section 4 - or, at least, to that version of twistor theory appropriate to asymptotically flat space-times [25]. The condition $\sigma^\circ = 0$ also seems to have some direct physical relevance to the definition of in- and out-states for zero rest-mass fields.

There is a method whereby it does appear to be possible to extract the Poincaré group in a reasonably canonical way, as a subgroup of the B.M.S. group, provided the space-time satisfies some very plausible additional requirements (near I°). Although no actual cross-section of I^+ may be shear-free, it is reasonable to expect that in the limit $u \to -\infty$ on I^+, such cross-sections (or at least, ones for which the "electric part" of σ° vanishes) will exist. Requiring that these limiting "shear-free cross-sections" be sent into one another, we can actually restrict the B.M.S. transformations to obtain a "canonically defined" subgroup of the B.M.S. group, which is isomorphic with the orthchronous Poincaré group [22].

The Poincaré group which emerges in this way may be thought of as that which has relevance to the outgoing field in the remote past, i.e. before any gravitational radiation has been emitted. In a similar way, we may extract another Poincaré subgroup of the B.M.S. group which has relevance to the remote future - after all the gravitational radiation has been emitted. (We simply replace the limit $u \to -\infty$ by $u \to +\infty$ in the above discussion.) There seems no reason to believe that these two Poincaré subgroups of the B.M.S. group will be the same, in general. The outgoing radiation which emerges between these limits could serve to "twist" one of these Poincaré subgroups in relation to the other. The asymptotic shear-free cross-sections of I^+ in the remote future would in general be "supertranslations" and not translations) of those in the remote past. There might, likewise, be a period of quiescence between bursts of outgoing radiation in which the system is near enough to being stationary, say, that yet another Poincaré subgroup of the B.M.S. group may be extracted. We would expect that the Poincaré group extracted in the quiescent period would, in general, be different yet again from that extracted in the remote past or the remote future. (Furthermore, the entire discussion could be repeated for I^- if desired.) Only the translation group would be common as a subgroup to all these copies of the Poincaré group, in the general case, (and the Lorentz group, common as a factor group). The role played by the B.M.S. group from the physical point of view would seem to lie in its relation to matters such as these.

To end this section, I shall be a little more explicit about how the gravitational radiation field is to be expressed, in terms of conformal infinity, and its relation to the asymptotic shear σ°. It will facilitate matters greatly to be able to employ the 2-component spinor formalism, so I shall assume the reader is familiar with this. If not, reference [28] should be consulted.

To begin with, consider the zero rest-mass free field equations for spin n/2. We can use a spinor $\phi_{AB\ldots L}$ with n indices to describe the field, this spinor being symmetric

$$\phi_{AB\ldots L} = \phi_{(AB\ldots L)} \tag{3.23}$$

and satisfying the free-field equation

$$\nabla^{AP'}\phi_{AB\ldots L} = 0 \tag{3.24}$$

where $\nabla_a = \nabla_{AA'}$ denotes* covariant derivative. When $n = 1$, this is the Weyl neutrino equation $\nabla^{AP'}\phi_A = 0$. When $n = 2$, it is Maxwell's free-field equations

$$\nabla_{[a}F_{bc]} = 0, \quad \nabla_b F^{ab} = 0$$

with

$$F_{ab} = \phi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\bar{\phi}_{A'B'} \, . \tag{3.25}$$

When $n = 4$, setting

$$K_{abcd} = \phi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\phi}_{A'B'C'D'} \tag{3.26}$$

we get a tensor with the symmetries of a Riemann tensor satisfying the Einstein vacuum equations

$$K_{abcd} = K_{[ab][cd]} = K_{cdab}\,, \quad K_{[abc]d} = 0, \quad K^a{}_{bad} = 0\,,$$

the differential equation (3.24) taking the form of Bianchi's identity

$$\nabla_{[a}K_{bc]de} = 0\,. \tag{3.27}$$

---

* Recall that the abstract index notation is being employed. The index label "a" stands for "AA'", so no connection symbols are required to connect tensor indices with spinor indices. This applies also to (3.25), (3.26) and later equations.

In flat space we can interpret solutions of (3.24) (or, equivalently (3.27)) as giving solutions of Einstein's vacuum equations in the weak field limit [29]. In curved space-time, the quantity K_{abcd} can be defined as before, but consistency requirements for (3.24) imply that K_{abcd} has to be related to the curvature tensor - actually to the Weyl conformal tensor defined by

$$C_{ab}{}^{cd} = R_{ab}{}^{cd} + 4\,P_{[a}{}^{[c}g_{b]}{}^{d]}$$

where P_{ab} is given by (2.13), so

$$C_{abcd} = C_{[ab][cd]} = C_{cdab}\ ,\quad C_{[abc]d} = 0\ ,\ C^{a}{}_{bad} = 0\ .$$

We can put

$$C_{abcd} = \Psi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\Psi}_{A'B'C'D'} \qquad (3.28)$$

where

$$\Psi_{ABCD} = \Psi_{(ABCD)}$$

the consistency requirement for (3.24) being

$$\Psi_{ABC(M}\phi^{ABC}{}_{D\ldots L)} = 0\ . \qquad (3.29)$$

If the space-time satisfies Einstein's vacuum equations (without cosmological term) then

$$C_{abcd} = R_{abcd}$$

so Bianchi's identity gives

$$\nabla_{[a}C_{bc]de} = 0$$

which is equivalent to

$$\nabla^{AP'}\Psi_{ABCD} = 0\ . \qquad (3.30)$$

A particular solution of (3.24) is, in this case

$$\phi_{ABCD} = \Psi_{ABCD} \qquad (3.31)$$

and (3.29) is then automatically satisfied. (For many vacuum space-times, the only solutions of (3.24) for $n = 4$ are those for which (3.31) is satisfied up to proportionality [30].) We may think of Ψ_{ABCD} as defining a spin 2 zero rest-mass field.

Let us now consider a conformal rescaling of the space-time metric according to

$$g_{ab} \to \hat{g}_{ab} = \Omega^2 g_{ab}\ ,\ g^{ab} \to \hat{g}^{ab} = \Omega^{-2}g^{ab}\ . \qquad (3.32)$$

We have

$$g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'}\ , \quad g^{ab} = \varepsilon^{AB}\varepsilon^{A'B'}$$

so we can take

$$\hat\varepsilon_{AB} = \Omega\varepsilon_{AB}\ , \quad \hat\varepsilon_{A'B'} = \Omega\varepsilon_{A'B'}\ , \quad \hat\varepsilon^{AB} = \Omega^{-1}\varepsilon^{AB}\ ,$$
$$\hat\varepsilon^{A'B'} = \Omega^{-1}\varepsilon^{A'B'} \tag{3.33}$$

The covariant derivative $\hat\nabla_a$ for the rescaled metric differs from the original one ∇_a, when acting on some spinor, by one term corresponding to each index of the spinor according to the scheme

$$\begin{aligned}\hat\nabla_{AA'}\chi_{H\ldots M'}^{B\ldots G'} = {}& \nabla_{AA'}\chi_{H\ldots M'}^{B\ldots G'} + \varepsilon_{A}{}^{B}T_{XA'}\chi_{H\ldots M'}^{X\ldots G'} + \ldots + \\ & + \varepsilon_{A'}{}^{G'}T_{AX'}\chi_{H\ldots M'}^{B\ldots X'} - T_{HA'}\chi_{A\ldots M'}^{B\ldots G'} - \ldots - \\ & - T_{AM'}\chi_{H\ldots A'}^{B\ldots G'}\ ,\end{aligned} \tag{3.34}$$

where

$$T_a = \Omega^{-1}\nabla_a\Omega = \nabla_a \log\Omega\ . \tag{3.35}$$

From this follows at once the invariance of the zero rest-mass equations (3.24) under conformal rescaling [6,8], where we set

$$\hat\phi_{AB\ldots L} = \Omega^{-1}\phi_{AB\ldots L}\ . \tag{3.36}$$

By (3.33), (3.25) and (3.26) we then have

$$\hat F_{ab} = F_{ab}\ , \quad \hat K_{abcd} = \Omega K_{abcd}\ .$$

If, for spin zero, we use the equation $(\nabla_a\nabla^a + \frac{1}{6}R)\phi = 0$ then again we have invariance with $\hat\phi = \Omega^{-1}\phi$.

Since zero rest-mass fields have this conformal invariance, we can refer them to the unphysical metric equally well as to the physical one, when discussing their properties. This allows us to consider the possibility of zero rest-mass fields defined on the whole of $\bar{M}$, where M is (weakly) asymptotically simple. If the field $\hat\phi_{AB\ldots L}$ is thus finite on I, we may re-interpret what asymptotic behaviour this entails in terms of the physical metric. It turns out [6] that such a physical field $\phi_{AB\ldots L}$ satisfies Sach's peeling-off property. Let ν be a null geodesic in M and choose a spinor dyad (or spin-frame) o^A, ι^A, normalized according to

$$o_A\iota^A = 1 \tag{3.37}$$

and parallelly propagated along ν, the spinor o^A having its null ("flagpole") direction pointing along ν. The affine parameter r along ν is scaled according to

$$o^A \bar{o}^{A'} \nabla_{AA'} r = 1 \ .$$

We define the components of $\phi_{AB\ldots L}$ with respect to o^A, ι^A as follows

$$\begin{aligned} \phi_0 &= \phi_{00\ldots0} = \phi_{AB\ldots L}\, o^A o^B \ldots o^L \\ \phi_1 &= \phi_{10\ldots0} = \phi_{AB\ldots L}\, \iota^A o^B \ldots o^L \\ &\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \phi_n &= \phi_{11\ldots1} = \phi_{AB\ldots L}\, \iota^A \iota^B \ldots \iota^L \end{aligned}$$

then the future peeling-off property states

$$\begin{aligned} \phi_0 &= \overset{0}{\phi}_0 r^{-n-1} + o(r^{-n-1}) \\ \phi_1 &= \overset{0}{\phi}_1 r^{-n} + o(r^{-n}) \\ &\ \vdots \qquad\quad \vdots \qquad\qquad \vdots \\ \phi_n &= \overset{0}{\phi}_n r^{-1} + o(r^{-1}) \ , \end{aligned} \tag{3.38}$$

for large positive r where $\overset{0}{\phi}_0, \overset{0}{\phi}_1, \ldots, \overset{0}{\phi}_n$ are constant on ν. The past peeling-off property is similar, but stated for large negative r.

The behaviour (3.38) is actually a direct consequence of the assumption that $\hat{\phi}_{AB\ldots L}$ is continuous on $\mathcal{I}^+$ (or $\mathcal{I}^-$), where we assume from now on that M is asymptotically flat, and may be derived immediately once it is established [6] that the rescaled dyad $\hat{o}^A$, $\hat{\iota}^A$, given by

$$\hat{o}_A = o_A \ , \quad \hat{\iota}_A = \Omega \iota_A \ , \quad \hat{o}^A = \Omega^{-1} o^A , \quad \hat{\iota}^A = \iota^A \ ,$$

is continuous at $\mathcal{I}$. We have, in fact

$$\begin{aligned} &\overset{0}{\phi}_0 = \hat{\phi}_{AB\ldots L} \hat{o}^A \hat{o}^B \ldots \hat{o}^L \ , \quad \overset{0}{\phi}_1 = \hat{\phi}_{AB\ldots L} \hat{\iota}^A \hat{o}^B \ldots \hat{o}^L , \ \ldots \ , \\ &\overset{0}{\phi}_n = \hat{\phi}_{AB\ldots L} \hat{\iota}^A \hat{\iota}^B \ldots \hat{\iota}^C \end{aligned}$$

at $\mathcal{I}^+$, where Ω is chosen so that

$$\Omega = r^{-1} + o(r^{-1}) \ . \tag{3.39}$$

In fact, (3.39) is equivalent to

$$\hat{\iota}_A \bar{\hat{\iota}}_{A'} = \hat{n}_a = -\hat{\nabla}_a \Omega$$

at $p = \nu \cap \mathscr{I}^+$, as in (2.14). In any case, the "flagpole" of $\hat{\iota}^A$ necessarily points in the null direction in $\mathscr{I}^+$.

The quantity ϕ^0_n, being the "1/r part of the field" along ν, measures the (outgoing) *radiation field* along ν. In fact, since ϕ^0_n does not involve $\hat{o}^A$, the radiation field is independent of the particular choice of null geodesic through p. Thus we may think of the outgoing radiation field as a function defined on $\mathscr{I}^+$, essentially dependent only on three parameters (u, θ, ϕ). In a similar way the incoming radiation field may be defined as the corresponding component defined on $\mathscr{I}^-$. For a *retarded* field, this incoming radiation field must *vanish*.

Suppose, now, that ν belongs to some congruence Γ of null geodesics. These could be the generators of the u = const. null hypersurfaces of a Bondi-type retarded time u, or they could constitute a more general system. We choose the spinor field o^A to have "flagpole" directions pointing along the null geodesics of the congruence. Then the *complex shear* σ and *complex divergence* ρ are defined by [24,8]

$$o^A o^B \nabla_{AA'} o_B = \sigma \bar{o}_{A'} \quad , \quad \bar{o}^{A'} o^B \nabla_{AA'} o_B = \rho o_A \ .$$

(Under $o^A \to \lambda o^A$, these rescale according to $\sigma \to \lambda^3 \bar{\lambda}^{-1} \sigma$, $\rho \to \lambda \bar{\lambda} \rho$.) The argument of σ defines the plane of maximum shear and the modulus of σ the magnitude of shear; the imaginary part of ρ measures the *rotation* of the congruence Γ and the real part its divergence. If the null geodesics generate null hypersurfaces, then $\rho = \bar{\rho}$.

Provided the null geodesics diverge into the future suitably (as they will if they generate Bondi-type hypersurfaces) we have

$$\sigma = \sigma^0 r^{-2} + O(r^{-3}) \quad , \quad \rho = -r^{-1} + O(r^{-2}) \tag{3.40}$$

for large positive r. (The situation in past directions is similar.) The quantity σ^0 in (3.40) is independent of r and defines the asymptotic shear of Γ, as was referred to earlier. In fact σ is a conformal density, with $\hat{\sigma} = \Omega^{-2}\sigma$. It follows that $\sigma^0 = \hat{\sigma}$ evaluated at $\mathscr{I}^+$.

Let us now assume that Einstein's vacuum equations hold near $\mathscr{I}^+$. We can adopt (3.31), and then the quantities

$$\Psi_0 = \phi_0 \ , \quad \Psi_1 = \phi_1 \ , \ \ldots \ , \ \Psi_4 = \phi_4 \tag{3.41}$$

measure the curvature. The (outgoing) gravitational radiation field is given by

$$\Psi^0_4 = \phi^0_4 \ .$$

It should be remarked that the spinor Ψ_{ABCD}, of which (3.41) are the components, describes the *conformal* curvature (cf. (3.28)) whether or not Einstein's vacuum equations hold. It is conformally *invariant*:

$$\hat{\Psi}_{ABCD} = \Psi_{ABCD} \; .$$

This should be contrasted with (3.36). With respect to the unphysical metric we obtain

$$\hat{\phi}_{ABCD} = \Omega^{-1}\hat{\Psi}_{ABCD} \; .$$

At first sight this would suggest that the assumption that $\hat{\phi}_{ABCD}$ is continuous at $\mathcal{I}^+$ is a very strong one. However, it can be shown [6] that the assumptions of Einstein's vacuum equations and asymptotic simplicity imply that $\hat{\Psi}_{ABCD} = 0$ on $\mathcal{I}^+$, and hence that $\hat{\phi}_{ABCD}$ *is* actually continuous at $\mathcal{I}^+$ as required. Incidentally, the B.M.S. group owes its existence to the (comparative) "flatness" of $\mathcal{I}^+$. This, in turn, is related to the vanishing of the conformal curvature at $\mathcal{I}^+$.

Taking Γ to be the system of generators of the u const. hypersurfaces in a Bondi-Sachs type coordinate system, we have the following formula [11,12,13] relating the asymptotic shear to the gravitational radiation field:

$$\frac{\partial^2\sigma^0}{\partial u^2} = -\bar{\Psi}_4^0 \; . \tag{3.42}$$

The time-integral of Ψ_4^0

$$N = -\frac{\partial\bar{\sigma}^0}{\partial u} \tag{3.43}$$

is essentially the Bondi-Sachs news function [11,12]. It can also be obtained [8,10] from the unphysical Ricci tensor R_{ab} on $\mathcal{I}^+$ by means of

$$\hat{R}_{AA'BB'}\hat{l}^A\hat{l}^B = 2N\,\bar{\hat{l}}_{A'}\bar{\hat{l}}_{B'} \; .$$

The quantity N enters into the Bondi-Sachs definition [11,12,10] of mass-momentum, evaluated at the retarded time u = const.

$$P^{\underline{a}} = \frac{1}{4\pi}\int W^{\underline{a}}(\sigma^0 N - \Psi_2^0)d\theta\,\sin\theta d\phi \tag{3.44}$$

where

$$W^0 = 1,\; W^1 = \sin\theta\cos\phi,\; W^2 = \sin\theta\sin\phi,\; W^3 = \cos\theta.$$

The rate of energy-momentum loss due to gravitational radiation is

$$\frac{\partial P^{\underline{a}}}{\partial u} = -\frac{1}{4\pi}\int W^{\underline{a}}N\bar{N}d\theta\,\sin\theta d\phi \; . \tag{3.45}$$

The intimate relation between the asymptotic shear and gravitational radiation should be clear from all this.

690 Relativistic symmetry groups. *Group theory in non-linear problems* (Lectures presented at the NATO Advanced Study Institute on mathematical physics, Istanbul, Turkey, August 7–18, 1972). Ed. A. O. Barut, pp. 1–58

4. TWISTOR THEORY

The theory of twistors [25,26,27] is a formalism for relativistic physics which affords a new approach to the description of quantized fields and to the treatment of space-time itself. In the twistor formalism, space-time points need not be employed as the primary objects in terms of which all else is to be expressed. Instead, the primary objects can be the twistors themselves.

The basic twistors form a complex four-dimensional vector space (considering the case of flat-space twistors only). This may be regarded as the vector space on which SU(2,2) matrices act, these matrices corresponding to conformal transformations of Minkowski space (compactified) according to the local isomorphism SU(2,2) → C(1,3). Such a twistor may be pictured in a classical physical way as a zero rest-mass particle with intrinsic spin. (There are three real parameters for the momentum, three for the spatial location, one for the degree of helicity, and one for a phase - which may be pictured in terms of a polarization plane. Thus there are eight real parameters in all, and they can be re-expressed to give four complex components defining a complex vector space.) Perhaps rather more accurate is to regard the twistor as a kind of "square-root" of the energy-momentum-angular-momentum structure of a zero-rest-mass particle. This is in a sense similar to the way that a two-component spinor may be regarded as a "square root" of a null vector. Space-time points may be interpreted, in twistor terms, as linear subspaces (of complex dimension two) of twistor space. This amounts to specifying a point in space-time, in terms of the totality of zero rest-mass particles which pass through that point.

The twistor formalism may also be adapted to curved space-times - especially those which are asymptotically flat. The twistor complex structure can be retained, but twistor space becomes, in a sense, curved (as a Kähler manifold). The complex numbers which describe twistor space are closely related to those more familiar complex numbers which play such a basic role in quantum theory. This points to a unification between quantum theory and space-time structure. The picture is also presented that in a quantized general relativity, instead of the null cones becoming "smeared" by the uncertainty principle (which has been a usual viewpoint), the points of space-time should, as a logical alternative, be themselves "smeared". (This amounts to the above-mentioned linear subspaces of twistor space being "smeared", according to the uncertainty principle.) But even in the classical theory the complex structure of twistor space plays a basic geometric role. This is in relation to the (asymptotic) shear-free condition on congruences of null geodesics that was referred to in Section 3. The twistor space complex structure is related also (via a contour integration) to the zero rest-mass free-field equations (3.24). This leads to a twistor formalism, involving complex contour integration, for the calculation of quantum scattering amplitudes and results in a form of massless quantum electrodynamics which appears to be free of the usual divergences.

I shall not discuss many of these matters in detail here. The interested reader is referred especially to ref. [25]. Only flat-space twistors will be discussed, and these primarily in relation to their role in the construction of different representations (finite or infinite) for the group SU(2,2), or its Lie algebra. Twistors seem to provide a very convenient means for expressing such representations generally. They also give an explicit realization of the local isomorphism SU(2,2) → C(1,3).

Let us consider the definition of a basic flat-space twistor. Suppose we are given a classical (finite) system in special relativity. Let P^a be the total 4-momentum of the system, and M^{ab} ($= -M^{ba}$) be its angular momentum with respect to some given space-time origin O. If we pass to a new origin Õ, whose position vector relative to O is X^a, then we have

$$P^a \to \tilde{P}^a = P^a \ , \quad M^{ab} \to \tilde{M}^{ab} = M^{ab} - 2X^{[a}P^{b]} \ . \tag{4.1}$$

The spin vector

$$S_a = \tfrac{1}{2}\, e_{abcd}\, P^b M^{cd} \tag{4.2}$$

is, like the momentum, displacement invariant:

$$S_a \to \tilde{S}_a = S_a \tag{4.3}$$

(The tensor e_{abcd} is the alternating tensor: $e_{abcd} = e_{[abcd]}$, with $e_{0123} = (-g)^{1/2}$ in a right-handed frame.) If the system is to be equivalent (in respect of its energy-momentum-angular-momentum structure) to a zero rest-mass particle, then we require

$$P_aP^a = 0, \ \ \underline{\text{with}}\ P^a\ \underline{\text{future-pointing}},\ \underline{\text{and}}\ S_a = sP_a \tag{4.4}$$

where s is the helicity, |s| being the spin. (If we choose units with $\hbar = 1$, then for quantum systems s takes half integer values.) The conditions (4.4) imply [25] that

$$P_{AA'} = \bar{\pi}_A\pi_{A'}, \quad M^{AA'BB'} = i\bar{\pi}^{(A}\omega^{B)}\varepsilon^{A'B'} - i\varepsilon^{AB}\pi^{(A'}\bar{\omega}^{B')} \tag{4.5}$$

for some $\pi_{A'}$, ω^A. The information contained in the two spinors ω^A, $\pi_{A'}$ is the same as that contained in P_a and M^{ab} except that the spinors contain the extra information of a "phase". That is to say, P_a and M^{ab} are invariant under

$$(\omega^A,\ \pi_{A'}) \to (e^{i\theta}\omega^A,\ e^{i\theta}\pi_{A'})$$

(θ real) but apart from this freedom, they uniquely define (and are uniquely defined by) ω^A and $\pi_{A'}$.

The two spinors ω^A, $\pi_{A'}$ serve to define a twistor Z^α

$$Z^\alpha \leftrightarrow (\omega^A, \pi_{A'}) \ . \tag{4.6}$$

The four complex components of Z^α are (with respect to the chosen origin 0 and given spin-frame)

$$Z^0 = \omega^0, \quad Z^1 = \omega^1, \ Z^2 = \pi_{0'}, \ Z^3 = \pi_{1'} \ . \tag{4.7}$$

Under change of origin $0 \to \tilde{0}$, we have

$$\omega^A \to \tilde{\omega}^A = \omega^A - iX^{AA'}\pi_{A'} \ , \quad \pi_{A'} \to \tilde{\pi}_{A'} = \pi_{A'} \ , \tag{4.8}$$

this being consistent with (4.1) and (4.5). The (abstract) twistor itself is unaffected by the change in origin, but the representation (4.7) of the twistor components undergoes the replacement (4.8) when 0 is replaced by $\tilde{0}$.

We can also view a twistor in a slightly different way. Let us regard $\tilde{0}$ as a variable point at which the spinor $\tilde{\omega}^A$ is defined. Thus $\tilde{\omega}^A$ gives us a spinor field. It is convenient to drop the tilde and write this field simply as ω^A, rewriting the first equation (4.8) as

$$\omega^A = \underset{\circ}{\omega}^A - ix^{AA'}\underset{\circ}{\pi}_{A'} \tag{4.9}$$

where the "$\circ$" placed under a spinor symbol means that this is a quantity defined at the origin, and where x^a is the position vector relative to this origin. The field (4.9) satisfies the equation

$$\nabla_{B'}^{(B}\omega^{A)} = 0 \tag{4.10}$$

and it is not hard to see that, conversely, the general solution of (4.10) is (4.9), for some (unique) $\underset{\circ}{\omega}^A$, $\underset{\circ}{\pi}_{A'}$. Thus, we can think of the twistor Z^α as being, in effect, a spinor field satisfying (4.10). This viewpoint has the virtue that the twistor concept is made conformally invariant. For it follows from (3.34) that (4.10) is conformally invariant (with $\hat{\omega}^A = \omega^A$) under the conformal rescaling (3.32). We can thus think of the field ω^A as being defined over the whole compactified Minkowski space. There is, however, a difficulty. It turns out that the field ω^A must jump by a factor i (or -i) as we cross the light cone at infinity. Thus we must, strictly speaking, regard the field ω^A as defined on the fourfold covering space of compactified Minkowski space (or, what amounts to the same thing, as a four-valued field on compactified Minkowski space). The reason for this four-valuedness is the (4-1) nature of the homomorphism SU(2,2) → C(1,3).

Suppose we choose a conformal rescaling $g_{ab} \to \hat{g}_{ab} = \Omega^2 g_{ab}$ which is of the type that renders the $\hat{g}_{ab}$ to be (like g_{ab}) flat. This means that Ω has a form proportional to

$$\Omega = \frac{Q_a Q^a}{(Q_b - x_b)(Q^b - x^b)} \tag{4.11}$$

(with Q^a constant). Since we can extract $\pi_{A'} (= \mathfrak{g}_{A'})$ from the field ω^A by

$$\nabla_{BB'}\omega^A = -i\varepsilon_B{}^A \pi_{B'} \tag{4.12}$$

[see (4.9)] either using the g_{ab} metric or the $\hat{g}_{ab}$ metric, we get (by (3.34))

$$\omega^A \rightarrow \hat{\omega}^A = \omega^A, \quad \pi_{A'} \rightarrow \hat{\pi}_{A'} = \pi_{A'} + iT_{AA'}\omega^A \tag{4.13}$$

under the conform rescaling (3.32), where

$$T_a = \nabla_a \log \Omega = \frac{2(Q_a - x_a)}{(Q_b - x_b)(Q^b - x^b)} . \tag{4.14}$$

Note that the transformations (4.8) and (4.13) are similar in form. They are each linear and unimodular in the components of the twistor Z^α (with (4.13) evaluated at the origin, so that it can be taken as referring to the components (4.7); also $\Omega = 1$ at the origin in (4.11), so the spin-frame normalization is unaffected, cf. (3.37)). Furthermore, they each leave invariant the expression

$$\begin{aligned} Z^\alpha \bar{Z}_\alpha &= \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^{A'} \\ &= 2\mathrm{Re}[\omega^0 \bar{\pi}_0 + \omega^1 \bar{\pi}_1] \end{aligned} \tag{4.15}$$

where the complex conjugate twistor to Z^α is defined by

$$\bar{Z}_\alpha \leftrightarrow (\bar{\pi}_A , \bar{\omega}^{A'}) . \tag{4.16}$$

The signature of the Hermitian form (4.15) is (++--), so that the transformations of components of Z^α given by (4.8) and (4.13) all belong to the group SU(2,2) whose invariant Hermitian form is $Z^\alpha \bar{Z}_\alpha$.

The same remark applies to the transformations of Z^α components induced by a Lorentz transformation (leaving the origin 0 invariant). For these are given by SL(2,C) transformations applied to ω^A, coupled with the corresponding dual conjugate SL(2,C) transformations applied to $\pi_{A'}$:

$$\omega^A \rightarrow \omega^B t_B{}^A , \quad \pi_{A'} \rightarrow \pi_{B'}(-\bar{t}^{B'}{}_{A'}), \tag{4.17}$$

the matrix of components of $t_B{}^A$ being unimodular (i.e., being a spin-matrix) so that $-t^B{}_A$ is inverse to $t_B{}^A$:

$$\varepsilon_{AB} t_C{}^A t_D{}^B = \varepsilon_{CD} , \quad \text{i.e.} \quad t_B{}^A(-t^C{}_A) = \varepsilon_B{}^C = (-t^A{}_B) t_A{}^C . \tag{4.18}$$

Clearly the transformation (4.17) is linear and unimodular on Z^α and preserves (4.15) (since (4.15) is a spin-scalar) as required. Furthermore, the dilations also belong to this SU(2,2). For expanding the Minkowski space by a factor k^2 $(k > 0)$ we have $x^a \to k^2x^a$ and $\Omega = k^2$ (= const.). This corresponds to

$$\omega^A \to k\omega^A \quad \text{and} \quad \pi_{A'} \to k^{-1}\pi_{A'} \tag{4.19}$$

at the origin in order that the relation between abstract spinors and vectors be preserved under $x^a \to k^2x^a$. (The conformal factor does not enter into (4.19) because of (4.13).) Choosing the spinor dyad o^A, ι^A as our spin-frame, with normalization

$$\varepsilon_{01} = \varepsilon_{AB}o^A\iota^B = o_A\iota^A = 1 \tag{4.20}$$

(cf. (3.37)), and using the fact that

$$\varepsilon_{AB} \to \Omega k^{-2}\varepsilon_{AB} = \varepsilon_{AB} \tag{4.21}$$

we get $o^A \to o^A$, $\iota^A \to \iota^A$, $o_A \to o_A$, $\iota_A \to \iota_A$ and the normalization (4.20) is preserved. Thus (4.19), when referred to this normalized spin-frame becomes

$$\omega^0 \to k\omega^0, \quad \omega^1 \to k\omega^1, \quad \pi_{0'} \to k^{-1}\pi_{0'}, \quad \pi_{1'} \to k^{-1}\pi_{1'} \tag{4.22}$$

(with $\pi_{0'} = \pi_{A'}\bar{o}^{A'}$, $\pi_{1'} = \pi_{A'}\bar{\iota}^{A'}$, $\omega^0 = \omega^A(-\iota_A)$, $\omega^1 = \omega^A o_A$). Clearly (4.22) is linear and homogeneous and preserves (4.15).

We have considered four types of transformation and seen that each of these induces a linear substitution of twistor components which belongs to the SU(2,2) for which (4.15) is the invariant Hermitian form. Three of these types of transformation (namely the translations, Lorenz transformations and dilations) clearly belong to C(1,3). In fact, the transformation (4.13), with (4.14), can also be regarded as belonging to C(1,3), but this requires a word of explanation. The transformation was given as simply a conformal rescaling from a flat space to a flat space. Notice that, by (4.11), Ω becomes infinite at $x^a = Q^a$. We can follow our conformal rescaling by a point mapping in which the light cone of $x^a = Q^a$ is sent to infinity and in which the metric regains the Minkowski form. This is given by

$$x^a \to \frac{x^a Q_c Q^c - Q^a x_c x^c}{(Q_b - x_b)(Q^b - x^b)} \tag{4.23}$$

These transformations form a 4-parameter subgroup of C(1,3) which is conjugate, with respect to inversion in the origin (cf. (1.20)) to the translation subgroup of C(1,3). The origin is transformed to itself under (4.23) and so is the tangent space at the origin.

Thus, the (active) C(1,3) transformation given by (4.23) is represented by an SU(2,2) transformation of twistor components, namely that obtained by taking the components of (4.13) at the origin. These four types of C(1,3) transformation will generate the whole of C(1,3). In fact, the infinitesmal transformations of these four types span the entire Lie algebra of C(1,3).

We can express the infinitesimal C(1,3) transformations as

$$x^a \to x^a + \varepsilon\xi^a \tag{4.24}$$

where ξ^a satisfies the conformal Killing equation

$$\nabla_{(a}\xi_{b)} = \tfrac{1}{4}\, g_{ab}(\nabla_c\xi^c) \tag{4.25}$$

(i.e., the Lie derivative of the metric with respect to ξ^a is a scalar multiple of itself). The general solution of (4.25) turns out to be

$$\xi_a = S_{ab}x^b + T_a + Kx_a + L_a(x_bx^b) - 2x_a(x_bL^b) \tag{4.26}$$

where S_{ab} is skew and each of S_{ab}, T_a, K, L_a is constant. The infinitesimal Lorentz transformations are described by S_{ab} (6 parameters), the translations by T_a (4 parameters), dilations by K (1 parameter), and special conformal transformations of the type (4.23) by L_a (4 parameters), giving 15 parameters in all. The twistor representations of these infinitesimal transformations are readily obtained from the foregoing discussion of the finite ones (cf. (4.17), (4.8), (4.22) and (4.13),respectively).

The invariant Hermitian form (4.15) has an interesting physical interpretation. By referring back to (4.2), (4.4) and (4.5) we can obtain

$$Z^\alpha\bar{Z}_\alpha = 2s\ . \tag{4.27}$$

The helicity (or spin) is thus a conformal invariant. When $s = 0$ we have

$$Z^\alpha\bar{Z}_\alpha = 0 \tag{4.28}$$

and Z^α (or $\bar{Z}_\alpha$) is referred to as a null twistor.

Null twistors can be pictured in a particularly simple way. Suppose, first, that $\pi_{A'}$ is not a multiple of $\bar{\omega}_{A'}$ (at the origin O), but that (4.28) holds. Choose $\tilde{O}$ to have position vector x^a relative to O and set

$$x^{AA'} = i(\bar{\omega}^{B'}\pi_{B'})^{-1}\omega^A\bar{\omega}^{A'} + h\bar{\pi}^A\pi^{A'}\ . \tag{4.29}$$

The vector x^a is real (h being real), the Hermitian nature of (4.29) following because $\bar{\omega}^{B'}\pi_{B'}$ is pure imaginary by (4.28) (see (4.15)). Substituting (4.29) into (4.8) we see that $\tilde{\omega}^A = 0$. Since h can be any real number, the locus of such points O is a *null straight line* Z in the direction of P^a. The spinor $\tilde{\omega}^A$ vanishes all along the null line Z and nowhere else (as one readily verifies). Thus, Z is the locus of points $\tilde{O}$, with respect to which the angular momentum $\tilde{M}^{ab}$ vanishes (see (4.5)). We can think of Z as the *world line* of the *zero rest-mass particle* to which the original classical system under consideration is equivalent. Knowledge of the momentum P^a and the location of Z (with Z parallel to P^a) will uniquely determine the energy-momentum-angular-momentum of the system. The twistor Z^α itself requires the one further piece of information given by the phase of $\pi_{A'}$. This null twistor may be pictured in terms of the world line Z together with the spinor $\pi_{A'}$, with the "flagpole" direction of $\pi_{A'}$ parallel to Z. The "flag plane" of $\pi_{A'}$ gives the remaining "phase" information, and can be pictured geometrically in terms of a kind of oriented "polarization plane" for the zero rest-mass particle.

We have considered only the case when $\pi_{A'}$ is not a multiple of $\bar{\omega}_{A'}$. If it *is* such a multiple, but $\pi_{A'} \neq 0$, the situation is no different (as is not hard to see) except that the null line Z lies in a null hyperplane through O and the specific form (4.29) cannot be used. However, if $\pi_{A'}$ *vanishes* (but still $\omega^A \neq 0$) then we do not get a finite line Z. Instead (as can be seen using a limiting argument or a conformal transformation), we obtain a null geodesic Z which generates the null cone at infinity (i.e. Z is a null geodesic through the point I at infinity). Again, the twistor may be pictured (actually up to a multiple ±1, ±i) as the null line Z, together with a spinor $\pi_{A'}$ defined at points of Z, the "flagpole" of $\pi_{A'}$ pointing along Z, where $\pi_{A'}$ is taken parallelly propagated along Z. This description is conformally invariant.

If X^α and Y^α are two null twistors corresponding, respectively to null lines X and Y, the necessary and sufficient condition for them to intersect (possibly at infinity - which would mean, when X and Y are finite lines, that they belong to the same null hyperplane, cf. Section 3) is twistor orthogonality:

$$X^\alpha \bar{Y}_\alpha = 0. \tag{4.30}$$

This is not hard to verify, using (4.8).

If the orthogonality condition (4.30) holds, then every twistor representing a null line through the intersection point R of X and Y will have the form

$$Z^\alpha = \lambda X^\alpha + \mu Y^\alpha . \tag{4.31}$$

(Clearly Z^α is null and orthogonal to each of X^α and Y^α.) Thus, the space-time point R may be represented in twistor space as the linear set (4.31) (a complex 2-plane through the origin $Z^\alpha = 0$ of twistor space).

We have considered the geometrical picture of a null twistor only, so far. If Z^α is non-null ($Z^\alpha \bar{Z}_\alpha \neq 0$) then there is no such simple picture. We cannot localize the zero rest-mass particle, representing our system, to a single world-line when the intrinsic spin is non-zero. A non-local description can be given, however, in terms of a certain twisting system of null straight lines known as a *Robinson congruence* [26]. The Robinson congruence associated with a given non-null twistor Z^α can be obtained either as the system of null lines represented by null twistors X^α satisfying

$$X^\alpha \bar{Z}_\alpha = 0$$

or, equivalently, as the set of integral curves of the vector field W^a, where $W^{AA'} = \omega^A \bar{\omega}^{A'}$, with ω^A being the solution of (4.10) which represents Z. (Incidentally, such vector fields are precisely the solutions of the conformal Killing equation (4.25) which are everywhere null.) The limiting case of a Robinson congruence which occurs when Z is *null*, is simply the system of null straight lines meeting the null line Z. If Z is also at infinity then this becomes a system of parallel null straight lines.

One property of a Robinson congruence is that it is everywhere *shear-free*. This property is a consequence of a theorem due to Kerr [26] which states that every congruence of null lines Z in Minkowski space which is defined by an equation

$$\Phi(Z^\alpha) = 0 ,$$

where Φ is holomorphic and homogeneous in Z^α, is *shear-free*. Conversely *every* shear-free null congruence in Minkowski space arises in this way - or else as a limiting case of such a construction. Thus the *holomorphic* nature of functions on twistor space is closely related to the shear-free condition which played such a role in Section 3. (We may think of the Cauchy-Riemann equations $\partial\Phi/\partial\bar{Z}_\alpha = 0$ as being essentially the shear-free condition.) There is also an analogue of this result for curved asymptotically flat space-times [23,25], the shear-free condition being taken *asymptotically* as $\sigma^0 = 0$.

Another role played by homogeneous holomorphic functions in twistor space is the generation of solutions of the zero rest-mass free-field equations (3.27) in flat space-time [25,26]. Let $f(W_\alpha)$ be holomorphic and homogeneous of degree -n-2 (n = 0,1,2,....) in W_α with suitable positioned singularities. Consider the contour integral

$$\phi_{AB\ldots L}(x^q) = \frac{1}{2\pi i} \oint \lambda_A \lambda_B \ldots \lambda_L \, f(\lambda_Q, -ix^{QQ'}\lambda_Q) \lambda^R d\lambda_R \qquad (4.32)$$

where $\lambda_A, \ldots, \lambda_L$ are n in number. The contour is a closed curve for each x^q, but can vary (continuously) with x^q. This defines a spinor field throughout Minkowski space (or part of Minkowski space - depending on the location of the singularities of f). The spinor field is symmetric, $\phi_{AB\ldots L} = \phi_{(AB\ldots L)}$, and (as follows at once upon differentiating (4.32) with respect to $x^{QQ'}$) satisfies the

CHAPTER 61

zero rest-mass free-field equations $\nabla^{AA'}\phi_{AB\ldots L} = 0$ (cf. (3.23) and (3.24)) automatically. The holomorphic (and homogeneous) nature of f is required in an essential way for this. In order to be able simply to differentiate under the integral sign, it is necessary that the value of the integral be unaltered if the path of integration is slightly deformed. We require, indeed, that (4.32) is a proper contour integral in this sense. If the contour is moved continuously over any region throughout which the integrand remains non-singular, then the value of the integral remains unaltered. It is only the way in which the contour "links" the singularity region (i.e., technically, the homology class to which the contour belongs, in the space over which f is non-singular) that is relevant in the position of the contour. And the condition that this be the case is that the exterior derivative [31] of the expression following the integral sign in (4.32) should vanish. This will be the case provided f is holomorphic (except at the specified regions of singularity) and homogeneous of degree -n-2.

There is another way of writing (4.32). Let us express the field $\phi_{AB\ldots L}(x^q)$ itself in twistor terms. The point defined by x^q may be specified by a two-dimensional linear subspace of twistor space. Equivalently, we may work in the dual twistor space and specify this point by the space spanned by two dual twistors U_α and V_α, given by

$$U_\alpha \leftrightarrow (\xi_A, -ix^{AA'}\xi_A),\ V_\alpha \leftrightarrow (\eta_A, -ix^{AA'}\eta_A). \qquad (4.33)$$

These twistors represent null lines U, V through the point defined by x^a (see Figure 13). For it follows from (4.8) that

$$Z^\alpha \leftrightarrow (ix^{AA'}\pi_{A'}, \pi_{A'})$$

if Z passes through the point with position vector x^a (i.e. $\tilde{\omega}^A = 0$ there). Taking the complex conjugate we get the form (4.33) (cf. (4.16)) provided the point is real. Without loss of generality we may assume ξ_A, η_A to be normalized thus:

$$\xi_A\eta^A = 1 \qquad (4.34)$$

We can express λ_A as a linear combination of ξ_A and η_A (keeping

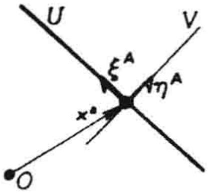


Fig. 13. The point with position vector x^a is the intersection of two null lines U, V. Twistors U^α and V^α, associated with these null lines, serve also to define a basis ξ^A, η^A for spin-vectors at that point.

ξ_A and η_A fixed for given x^a). Since the integrand in (4.32) is invariant under $\lambda_A \to \mu\lambda_A$ (where μ can be different at different points of the contour) - because of the homogeneity degree of f - we can put

$$\lambda_A = \lambda\xi_A + \eta_A \, .$$

Then $\xi_A\lambda^A = 1$ and $\lambda^R d\lambda_R = d\lambda$, whence

$$\xi_A\xi_B\dots\xi_L\phi^{AB\dots L}(x^q) = \frac{1}{2\pi i}\oint f(\lambda U_\alpha + V_\alpha)d\lambda = \phi(V_\alpha, U_\alpha) \qquad (4.35)$$

say. We can use the integral (4.35) to define $\phi(V_\alpha, U_\alpha)$ whether or not the normalization (4.34) is maintained. The result is a function ϕ which is homogeneous of degree $(-n-1)$ in V_α and homogeneous of degree -1 in U^α. This defines $\phi^{AB\dots L}\xi_A\xi_B\dots\xi_L$ at each point and for each ξ_A. Thus the field $\phi^{AB\dots L}$, being symmetric, is defined by the twistor function ϕ. From (4.35) we derive the property

$$\phi(V_\alpha + kU_\alpha,\ U_\alpha) = \phi(V_\alpha,\ U_\alpha) \qquad (4.36)$$

where k is arbitrary, with $|k|$ not too large. (This corresponds to the fact that (4.35) is independent of η_A.) Equation (4.36) can be restated as the property that ϕ can be expressed as a function of $V_{[\alpha}U_{\beta]}$ and U_α:

$$\phi(V_\alpha,\ U_\alpha) = F(V_{[\alpha}U_{\beta]},\ U_\alpha) \qquad (4.37)$$

(cf. ref. [35]). The significance of this is that $V_{[\alpha}U_{\beta]}$ effectively stands for the linear space spanned by V_α and U_α, so it represents the point with position vector x^a. The role of V is that it merely singles out this particular point on the line U. The direction of U has significance (provided $n \neq 0$) but not the direction of V. Another way of writing (4.36) (or (4.37)) may be derived:

$$U_\alpha \frac{\partial\phi}{\partial V_\alpha} = 0. \qquad (4.38)$$

(Differentiate (4.36) with respect to k.) This equation may be compared with those (Euler's theorem) which state the homogeneity degrees of ϕ:

$$U_\alpha \frac{\partial\phi}{\partial U_\alpha} = -\phi \ , \quad V_\alpha \frac{\partial\phi}{\partial V_\alpha} = (-n-1)\phi. \qquad (4.39)$$

Functions of twistors, which satisfy equations like (4.38) and (4.39) have an especial significance in the representation theory of U(2,2). Let us revert to twistor space, rather than the dual space, and consider, more generally, a holomorphic function Ψ of four twistors W^α, X^α, Y^α, Z^α which satisfies

$$X^\alpha \frac{\partial\Psi}{\partial W^\alpha} = Y^\alpha \frac{\partial\Psi}{\partial W^\alpha} = Z^\alpha \frac{\partial\Psi}{\partial W^\alpha} = 0,\ Y^\alpha \frac{\partial\Psi}{\partial X^\alpha} = Z^\alpha \frac{\partial\Psi}{\partial X^\alpha} = 0,\ Z^\alpha \frac{\partial\Psi}{\partial Y^\alpha} = 0 \qquad (4.40)$$

and

$$W^{\alpha}\frac{\partial\Psi}{\partial W^{\alpha}} = w\Psi,\ X^{\alpha}\frac{\partial\Psi}{\partial X^{\alpha}} = x\Psi,\ Y^{\alpha}\frac{\partial\Psi}{\partial Y^{\alpha}} = y\Psi,\ Z^{\alpha}\frac{\partial\Psi}{\partial Z^{\alpha}} = z\Psi. \qquad (4.41)$$

By (4.41) Ψ is homogeneous in W^{α}, X^{α}, Y^{α}, Z^{α} of respective degrees w,x,y,z. By (4.40) we have

$$\Psi(W^{\alpha}+aX^{\alpha}+bY^{\alpha}+cZ^{\alpha},\ X^{\alpha}+pY^{\alpha}+qZ^{\alpha},\ Y^{\alpha}+tZ^{\alpha},\ Z^{\alpha}) = \Psi(W^{\alpha},\ X^{\alpha},\ Y^{\alpha},\ Z^{\alpha}) \qquad (4.42)$$

(for small enough $|a|,\ldots,|t|$) since the partial derivatives with respect to $a,b,\ldots,t$ all vanish and consequently Ψ can be re-expressed in terms of

$$W^{[\alpha}X^{\beta}Y^{\gamma}Z^{\delta]},\ X^{[\alpha}Y^{\beta}Z^{\gamma]},\ Y^{[\alpha}Z^{\beta]},\ Z^{\alpha}$$

(it follows that the dependence of Ψ on W^{α} is essentially trivial, Ψ being Δ^{-w} times a function of X^{α}, Y^{α}, Z^{α} only, Δ being as in (4.44).)

In fact, any such function can also be referred to the dual twistor space. Introduce

$$\begin{aligned} \tilde{W}_{\delta} &= \Delta\varepsilon_{\alpha\beta\gamma\delta}\,W^{\alpha}X^{\beta}Y^{\gamma}, & \tilde{X}_{\gamma} &= \Delta\varepsilon_{\alpha\beta\gamma\delta}\,W^{\alpha}X^{\beta}Z^{\delta},\\ \tilde{Y}_{\beta} &= \Delta\varepsilon_{\alpha\beta\gamma\delta}\,W^{\alpha}Y^{\gamma}Z^{\delta}, & \tilde{Z}_{\alpha} &= \Delta\varepsilon_{\alpha\beta\gamma\delta}\,X^{\beta}Y^{\gamma}Z^{\delta}, \end{aligned} \qquad (4.43)$$

where

$$\Delta = (\varepsilon_{\alpha\beta\gamma\delta}\,W^{\alpha}X^{\beta}Y^{\gamma}Z^{\delta})^{-1} \qquad (4.44)$$

and where $\varepsilon_{\alpha\beta\gamma\delta}$ is the alternating twistor ($\varepsilon_{\alpha\beta\gamma\delta} = \varepsilon_{[\alpha\beta\gamma\delta]}$, $\varepsilon_{0123} = 1$). Setting $\tilde{\Delta} = \Delta^{-1}$ and $\tilde{\varepsilon}^{\alpha\beta\gamma\delta} = \varepsilon^{\alpha\beta\gamma\delta}$ ($=\varepsilon^{[\alpha\beta\gamma\delta]}$, with $\varepsilon^{0123} = 1$), we obtain expressions for W^{δ}, X^{γ}, Y^{β}, Z^{α} which are simply the "tilde versions" of (4.43). Thus we can re-express Ψ in terms of $\tilde{W}_{\alpha}$, etc. to obtain

$$\Psi(W^{\alpha},\ X^{\alpha},\ Y^{\alpha},\ Z^{\alpha}) = \chi(\tilde{W}_{\alpha},\ \tilde{X}_{\alpha},\ \tilde{Y}_{\alpha},\ \tilde{Z}_{\alpha}).$$

The function χ satisfies conditions analogous to (4.40), (4.41) and (4.42), the homogeneity degrees being $\tilde{w} = -z$, $\tilde{x} = -y$, $\tilde{y} = -x$, $\tilde{z} = -w$. For the case when $\tilde{W}_{\alpha} = \bar{W}_{\alpha}$, $\tilde{X}_{\alpha} = \bar{X}_{\alpha}$, $\tilde{Y}_{\alpha} = \bar{Y}_{\alpha}$, $\tilde{Z}_{\alpha} = \bar{Z}_{\alpha}$, such functions can be given rather direct space-time interpretations, but this matter will not be entered into here [34,35].

Functions of this type form a representation space for the Lie algebra of U(2,2) (and in some cases for the group itself) which, for fixed, w,x,y,z are eigenstates of the four Casimir operators. If w,x,y,z are non-negative integers, then the space of Ψ's contains as a subspace the space of polynomials in W^{α}, X^{α}, Y^{α}, Z^{α} homogeneous of respective degrees w,x,y,z. This subspace is finite-dimensional and in fact forms an irreducible representation space for the group U(2,2). We can write such Ψ's

$$\Psi(W^{\alpha},\ldots Z^{\alpha}) = \Psi_{\alpha_1\ldots\alpha_w\beta_1\ldots\beta_x\gamma_1\ldots\gamma_y\delta_1\ldots\delta_z} W^{\alpha_1}\ldots W^{\alpha_w}\ldots Z^{\delta_1}\ldots Z^{\delta_z}$$

where $\Psi_{\ldots.}$ has "Young tableau symmetry" corresponding to the partition $\{z,y,x,w\}$ (with $w \leqslant x \leqslant y \leqslant z$). That is,

$$\Psi_{\alpha_1\ldots\beta_1\ldots\gamma_1\ldots\delta_1\ldots} = \Psi_{(\alpha_1\ldots)(\beta_1\ldots)(\gamma_1\ldots)(\delta_1\ldots)}$$

and

$$\Psi_{(\alpha_1\ldots\alpha_w\beta_1)\beta_2\ldots.} = 0,\ \ldots,$$

$$\Psi_{(\alpha_1\ldots\alpha_w|\beta_1\ldots.\gamma_z|\delta_1)\delta_2\ldots\delta_z} = 0,\ \ldots,$$

$$\Psi_{\alpha_1\ldots..(\gamma_1\ldots\gamma_y\delta_1)\delta_2\ldots\delta_z} = 0\ .$$

(Indices between vertical bars are excluded from symmetrization.) The most general finite-dimensional irreducible representation of U(2,2) can in fact be put in this form. However, it is clear that non-polynomial Ψ's also exist for non-negative w,x,y,z, so the complete space of Ψ's is not irreducible. (A similar situation occurs for representations of SL(2,C) [32,33].) For other values of w,x,y,z the complete Ψ-space may be irreducible.

Let us see, in the general case, how the space of Ψ's may be regarded as a representation space of U(2,2) or SU(2,2) or of their Lie algebras. The group U(2,2) may be regarded as the group of linear transformations of twistor space to itself which preserves the Hermitian form $Z^{\alpha}\bar{Z}_{\alpha}$. These are given by

$$Z^{\alpha} \to t^{\alpha}{}_{\beta}\, Z^{\beta}$$

where

$$t^{\alpha}{}_{\beta}\bar{t}_{\gamma}{}^{\beta} = \delta^{\alpha}_{\gamma} \tag{4.45}$$

the twistor complex conjugate of $t^{\alpha}{}_{\beta}$ being $\bar{t}_{\alpha}{}^{\beta}$, so that in terms of components we have

$$\begin{pmatrix} \bar{t}_0{}^0 & \bar{t}_0{}^1 & \bar{t}_0{}^2 & \bar{t}_0{}^3 \\ \bar{t}_1{}^0 & \bar{t}_1{}^1 & \bar{t}_1{}^2 & \bar{t}_1{}^3 \\ \bar{t}_2{}^0 & \bar{t}_2{}^1 & \bar{t}_2{}^2 & \bar{t}_2{}^3 \\ \bar{t}_3{}^0 & \bar{t}_3{}^1 & \bar{t}_3{}^2 & \bar{t}_3{}^3 \end{pmatrix} = \begin{array}{c}\text{complex}\\ \text{conjugate}\\ \text{of}\end{array} \begin{pmatrix} t^2{}_2 & t^2{}_3 & t^2{}_0 & t^2{}_1 \\ t^3{}_2 & t^3{}_3 & t^3{}_0 & t^3{}_1 \\ t^0{}_2 & t^0{}_3 & t^0{}_0 & t^0{}_1 \\ t^1{}_2 & t^1{}_3 & t^1{}_0 & t^1{}_1 \end{pmatrix} \tag{4.46}$$

To obtain SU(2,2) the matrices are taken to be unimodular; equivalently the twistor $\varepsilon^{\alpha\beta\gamma\delta}$ is invariant in the sense that

$$t^{\alpha}{}_{\kappa}t^{\beta}{}_{\lambda}t^{\gamma}{}_{\mu}t^{\delta}{}_{\nu}\varepsilon^{\kappa\lambda\mu\nu} = \varepsilon^{\alpha\beta\gamma\delta}\ .$$

CHAPTER 61

Each $t^\alpha{}_\beta$ in U(2,2) is represented, in the space of functions ψ, by the linear transformation of such functions:

$$\psi(W^\alpha, X^\alpha, Y^\alpha, Z^\alpha) \to \psi(t^\alpha{}_\beta W^\beta, t^\alpha{}_\beta X^\beta, t^\alpha{}_\beta Y^\beta, t^\alpha{}_\beta Z^\beta). \qquad (4.47)$$

(This is clearly linear because $a\psi+b\phi \to a\psi'+b\phi'$, where $\psi \to \psi'$, $\phi \to \phi'$.) However ψ will in general possess singularities, the transform of ψ possessing singularities in different regions. There is thus no guarantee that a space of such functions can be constructed which are non-singular over some definite fixed region. If we consider infinitesimal transformations, on the other hand, this problem does not arise since the singularity regions remain fixed.

Putting

$$t^\alpha{}_\beta = \delta^\alpha{}_\beta + i\varepsilon h^\alpha{}_\beta$$

where ε is infinitesimal we have

$$h^\alpha{}_\beta = \bar{h}_\beta{}^\alpha$$

by (4.45), (4.46) where, in addition

$$h^\alpha{}_\alpha = 0$$

if $t^\alpha{}_\beta$ belongs to SU(2,2). The transformation (4.47) now becomes

$$\psi \to \psi + i\varepsilon h^\alpha{}_\beta L^\beta{}_\alpha \psi$$

where

$$L^\beta{}_\alpha = W^\beta \frac{\partial}{\partial W^\alpha} + X^\beta \frac{\partial}{\partial X^\alpha} + Y^\beta \frac{\partial}{\partial Y^\alpha} + Z^\beta \frac{\partial}{\partial Z^\alpha}\ .$$

A complete set of Casimir operators for U(2,2) is

$$L^\alpha{}_\alpha,\ L^\alpha{}_\beta L^\beta{}_\alpha,\ L^\alpha{}_\beta L^\beta{}_\gamma L^\gamma{}_\alpha,\ L^\alpha{}_\beta L^\beta{}_\gamma L^\gamma{}_\delta L^\delta{}_\alpha \qquad (4.48)$$

or, equivalently,

$$K_1 = K_\alpha{}^\alpha,\ K_2 = K_\beta{}^\alpha K_\alpha{}^\beta,\ K_3 = K_\gamma{}^\beta K_\beta{}^\alpha K_\alpha{}^\gamma,$$

$$K_4 = K_\delta{}^\gamma K_\gamma{}^\beta K_\beta{}^\alpha K_\alpha{}^\delta \qquad (4.49)$$

where

$$K_\beta{}^\alpha = L^\alpha{}_\beta + 4\,\delta^\alpha_\beta\ .$$

To obtain the Casimir operators for SU(2,2) we replace $K^\alpha{}_\beta$ or $L^\alpha{}_\beta$ by its trace-free part, in (4.48) or (4.49). The results are certain polynomials in $K_1, \ldots, K_4$, but I shall not bother with them here.

A. Qadir [34,35] has shown that when the operators act on a ψ satisfying (4.40) and (4.41), then

$$\begin{aligned} K_1 &= 10 + S_1, \quad K_2 = 20 + 5S_1 + S_2, \\ K_3 &= 35 + 15S_1 + \frac{1}{2} S_1{}^2 + \frac{11}{2} S_2 + S_3 \\ K_4 &= 36 + 35S_1 + 3S_1{}^2 + \frac{1}{3} S_1{}^3 + 18S_2 + \frac{20}{3} S_3 + S_4 \qquad (4.50) \end{aligned}$$

where

$$\begin{aligned} S_1 &= \omega + \xi + \eta + \zeta, \quad S_2 = \omega^2 + \xi^2 + \eta^2 + \zeta^2, \\ S_3 &= \omega^3 + \xi^3 + \eta^3 + \zeta^3, \quad S_4 = \omega^4 + \xi^4 + \eta^4 + \zeta^4, \qquad (4.51) \end{aligned}$$

the numbers of homogeneity ω,ξ,η,ζ being defined by

$$\omega = \omega, \ \xi = x+1. \ \eta = y+2, \ \zeta = z+3 .$$

This curious, but no doubt significant, fact that the eigenvalues of the Casimir operators are symmetric functions of the numbers of homogeneity [34,35], is related to the existence of certain contour integral expressions of which (4.35) is a special case. Consider (4.35) first and re-express it

$$\phi(Y^\alpha, Z^\alpha) = \oint \psi(\lambda Z^\alpha + Y^\alpha) d\lambda \ , \qquad (4.52)$$

$\psi(Z^\alpha)$ being homogeneous of degree z. Now ϕ satisfies (4.40) and (4.41) (or (4.42)) as does ψ (trivially), by (4.38), (4.39). The homogeneity degrees of ψ in Y^α and in Z^α are z+1 and -1, respectively. The numbers of homogeneity $\{\omega,\xi,\eta,\zeta\}$ are $\{0,1,2,z+3\}$ in the case of ψ, and are $\{0,1,z+3.2\}$ in the case of ϕ. Note that these numbers are merely permuted in the passage from ψ to ϕ. Thus the values of the Casimir operators are the same for ϕ as for ψ. In fact this can be seen more directly from (4.52) [35].

The expression (4.52) has been generalized by Qadir in the following way [35]. Let $\psi(W^\alpha,X^\alpha,Y^\alpha,Z^\alpha)$ be as in (4.40), (4.41) and put

$$\begin{aligned} \rho(W^\alpha,X^\alpha,Y^\alpha, Z^\alpha) &= \oint \psi(Z^\alpha, W^\alpha+\lambda Z^\alpha, X^\alpha+\mu Z^\alpha, Y^\alpha+\nu Z^\alpha) d\lambda\wedge d\mu\wedge d\nu, \\ \sigma(W^\alpha,X^\alpha,Y^\alpha,Z^\alpha) &= \oint \psi(W^\alpha,Z^\alpha,X^\alpha+\mu Z^\alpha,Y^\alpha+\nu Z^\alpha) d\mu\wedge d\nu, \qquad (4.53) \\ \tau(W^\alpha,X^\alpha,Y^\alpha,Z^\alpha) &= \oint \psi(W^\alpha,X^\alpha,Z^\alpha,Y^\alpha+\nu Z^\alpha) d\nu. \end{aligned}$$

the contours and singularity structure of ψ being suitably arranged. It can be verified that ρ,σ,τ are all functions of the required type, the numbers of homogeneity being $\{\zeta,\omega,\xi,\eta\}$, $\{\omega,\zeta,\xi,\eta\}$, $\{\omega,\xi,\zeta,\eta\}$. Since an arbitrary permutation of $\{\omega,\xi,\eta,\zeta\}$ can be built up by successive applications of permutations of these types (or even of the first two) and since, as can be simply shown, ρ, σ and τ must have the same eigenvalues of any Casimir operator as ψ, the invariance of the Casimir operators under permutation of the numbers of homogeneity follows at once. Thus these eigenvalues must be expressible in terms of (4.51) as was done in (4.50). It would be interesting to know whether the integrals (4.53) can, like (4.52), find important application within twistor theory itself. For a different kind of application of contour integration in twistor theory which relates to quantum scattering theory see ref. [25].

I have not attempted here to connect this twistor method of constructing representations of U(2,2) or SU(2,2) with previously existing work - for which the reader is referred, especially, to ref. [36]. In particular, the question of unitarity of the representations has not been discussed here. I feel that there is much scope for further study of the interrelations between the twistor formalism and other approaches.

ACKNOWLEDGEMENTS

I am grateful to Boston University for their hospitality and for providing typing facilities while this article was in preparation. I would also like to thank Judith Daniels for suggesting several improvements to the manuscript.

REFERENCES

1. Cartan, M.E., Ann. Sci. Ecole Normale Superieure, 31 (1914) 263.
2. H. Weyl, The Classical Groups, Princeton University Press, Princeton, New Jersey, 1939.
3. N.H. Kuiper, Ann. Math. 50 916 (1949).
4. H. Rudberg, Disseration, University of Uppsala, Sweden, 1958.
5. R. Penrose, In Proc. 1962 Conf. on Relativistic Theories of Gravitation, Warsaw, Polish Acad. Sci., Warsaw, 1965.
6. R. Penrose, Proc. Roy. Soc. (Lond.) A 284 159 (1965).
7. L.P. Eisenhart, Continuous Groups of Transformations, Princeton University Press 1933; reprinted by Dover 1961.
8. R. Penrose, in Battelle Rencontres 1967, C.M. DeWitt & J.A. Wheeler (eds.), Benjamin, New York, 1968.
9. J.M. Bardeen and W.H. Press, J. Math. Phys. 14 7 (1973).
10. R. Penrose, in Relativity, Groups and Topology, C.M. DeWitt & B.S. DeWitt (eds.), Les Houches Lectures 1963, Blackie & Son, London, 1964.
11. H. Bondi, A.W.K. Metzner and M.J.G. van der Burg. Proc. Roy. Soc.(London) A 269 21 (1962).
12. R.K. Sachs, Proc. Roy. Soc. (London) A 270 103 (1962).
13. E.T. Newman and T.W.J. Unti, J. Math. Phys. 3 891 (1962).
14. R.K. Sachs, Phys. Rev. 128 2851 (1962).

15. V.J. Cantoni, *J. Math. Phys.* **7** 1361 (1966); **8** 1700 (1967).
16. V.J. Cantoni, *Accad. Naz. Linc.* (13) **43** 30 (1967).
17. P.J. McCarthy, *Proc. Roy. Soc. (London)* **A 330** 517 (1972).
18. P.J. McCarthy, *J. Mat. Phys.* **13** 1837 (1972).
19. P.J. McCarthy, *Proc. Roy. Soc. (London)* **A 333**, 317 (1973).
20. M. Crampin and P.J. McCarthy, Representations of the B.M.S. Group III (to appear).
21. R. Geroch and E.T. Newman, *J. Math. Phys.* **12** 314 (1971).
22. E.T. Newman and R. Penrose, *J. Math. Phys.* **7** 863 (1966).
23. B. Aronson, R. Lind, J. Messmer and E.T. Newman, *J. Math. Phys.* **12** 2462 (1971).
24. E.T. Newman and R. Penrose, *J. Math. Phys.* **3** 566 (1962); **4** 998 (1963).
25. R. Penrose and M.A.H. MacCallum, *Physics Reports* **6C** 242 (1973)
26. R. Penrose, *J. Math. Phys.* **8** 345 (1967); **10** 38 (1969).
27. R. Penrose, *Int. J. Theor. Phys.* **1** 61 (1968).
28. F.A.E. Pirani, in *Lectures on General Relativity*, Brandeis. Summer Institute in Theoretical Physics 1964, Vol. 1, Prentice Hall, Englewood Cliffs, New Jersey, 1965.
29. R.K. Sachs and P.G. Bergmann, *Phys. Rev.* **112** 674 (1958).
30. P. Bell and P. Szekeres, *Int. J. Theor. Phys.* **6** 111 (1972).
31. H. Flanders, *Differential Forms*, Academic Press, New York, 1963.
32. M.A. Naimark, *Linear Representations of the Lorentz Group*, Pergamon, Oxford, 1964.
33. I.M. Gel'fand, M.I. Graev and N.Ya. Vilenkin, *Generalized Functions*, Vol. 5, Academic Press, New York, 1966.
34. A. Qadir, Ph.D. Thesis, London University, 1971.
35. A. Qadir, Trieste preprint 1972.
36. T. Yao, *J. Math. Phys.* **8** 1931 (1967).

62

GRAVITATIONAL COLLAPSE

Lecture notes for the International School of Cosmology and Gravitation at the 'Ettore Majorana' Centre for Scientific Culture, Erice, Italy, 1974

Singularity Theorem

(Hawking & Penrose 1970)

A space-time which

(i) contains no closed timelike curves

(ii) satisfies Einstein's equations (without cosmological term) and the energy condition $\rho + p_i \geq 0$, $\rho + \sum_i p_i \geq 0$

(iii) is sufficiently general (i.e. $t_{[a}R_{b]cd[e}t_{f]}t^c t^d \neq 0$ somewhere along each timelike or null geodesic, t^a being the tangent vector)

and

(iv) contains either

(a) a closed spacelice hypersurface

or (b) a trapped surface

or (c) a point p for which the divergence of the null geodesics through p changes sign somewhere to the past of p

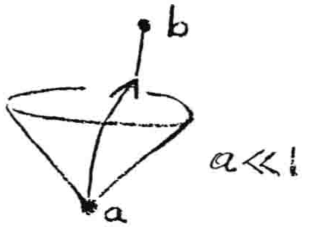
cannot be geodesically complete for all timelike and null geodesics.

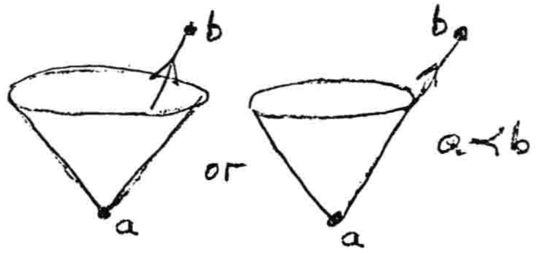
Causality Relations

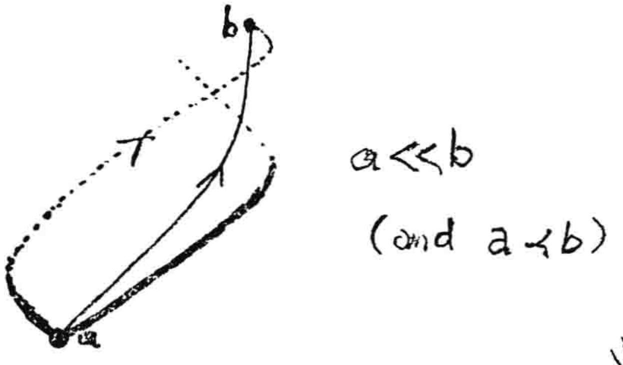
$a \ll b$: there exists a future-timelike curve from a to b

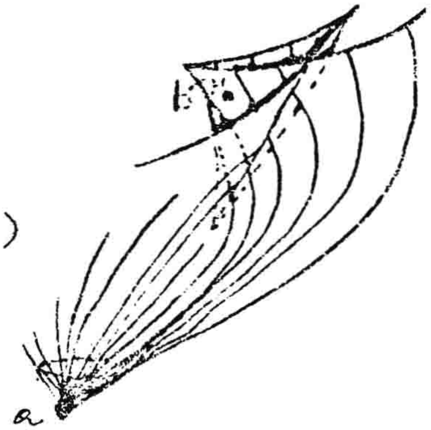
$a \prec b$: there exists a future-causal curve from a to b or $a = b$

(note: "causal" = "timelike or null"
= "non-spacelike")







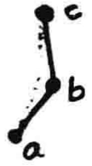


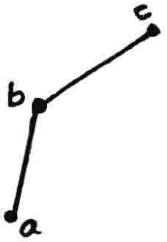
$$a \ll b \quad \& \quad b \ll c \quad \Rightarrow \quad a \ll c$$

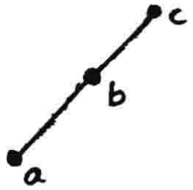
$$a \ll b \quad \& \quad b \prec c \quad \Rightarrow \quad a \ll c$$

$$a \prec b \quad \& \quad b \ll c \quad \Rightarrow \quad a \ll c$$

$$a \prec b \quad \& \quad b \prec c \quad \Rightarrow \quad a \prec c$$



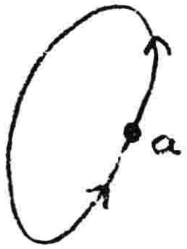




$$a \prec a$$

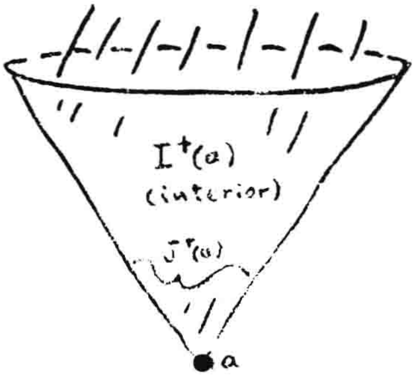
$a \ll a \Rightarrow \exists$ closed timelike curves

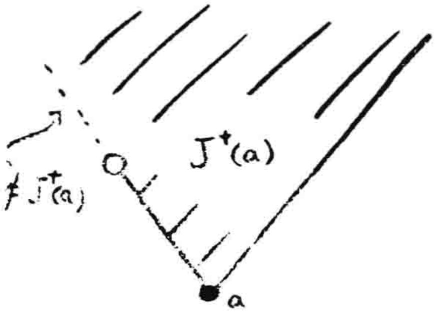
$a \prec b \ \& \ b \prec a \ \& \ a \neq b \Rightarrow \exists$ closed null curves



$$I^+(a) = \{x \mid a \ll x\}, \quad I^-(a) = \{x \mid x \ll a\}$$

$$J^+(a) = \{x \mid a \prec x\}, \quad J^-(a) = \{x \mid x \prec a\}$$

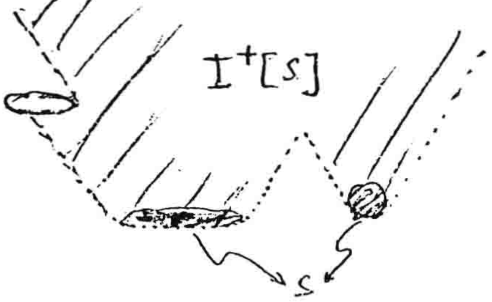

I⁺(a)
(interior)
J⁺(a)
a

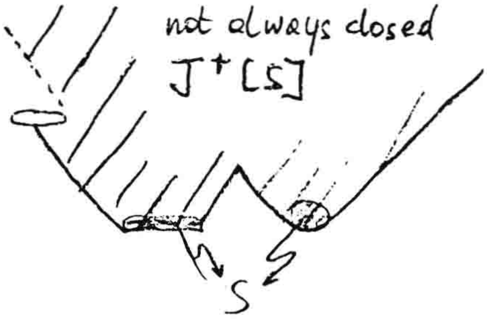


$$I^+[S] = \bigcup_{x\in S} I^+(x), \quad I^-[S] = \bigcup_{x\in S} I^-(x)$$

$$J^+[S] = \bigcup_{x\in S} J^+(x), \quad J^-[S] = \bigcup_{x\in S} J^-(x)$$

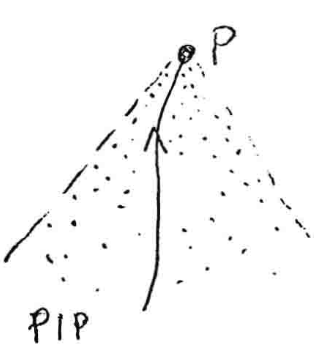
any $S \subset M$ M = space-time

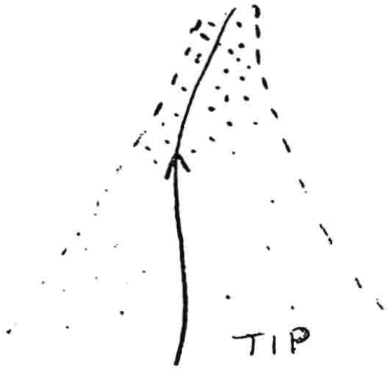



not always closed
J⁺[S]
S

γ = timelike (or causal) curve

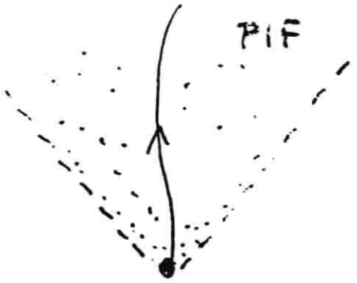
$$I^-[\gamma] = \text{"IP"} \begin{cases} \text{PIP} : & I^-[\gamma] = I^-(p) \\ \text{TIP} : & I^-[\gamma] \neq I^-(p) \\ & \text{(any } p\text{)} \end{cases}$$

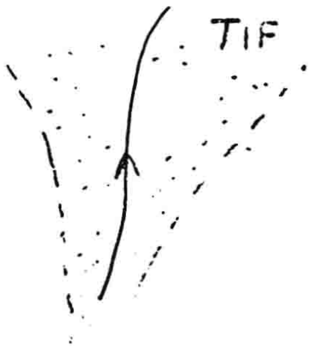




$\boxed{\hat{M} = \text{set of IPs}}$

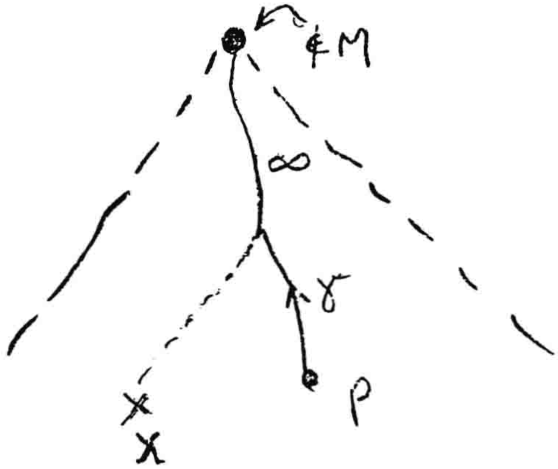
$$I^+[\gamma] = \text{"IF"} \begin{cases} \text{PIF} : & I^+[\gamma] = I^+(p) \\ \text{TIF} : & I^-[\gamma] \neq I^+(p) \\ & \text{(any } p\text{)} \end{cases}$$





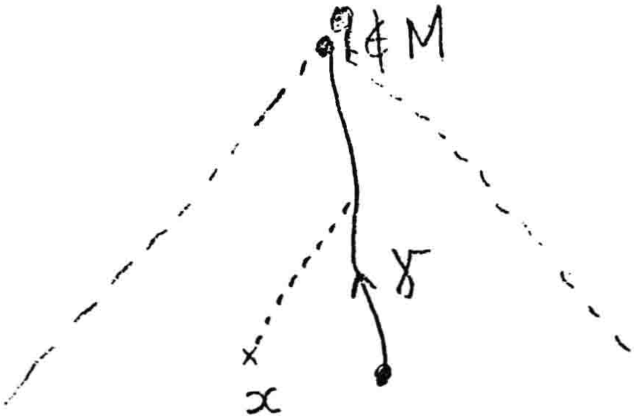
$\boxed{\check{M} = \text{set of IFs}}$

Point at (future) infinity : ∞TIP

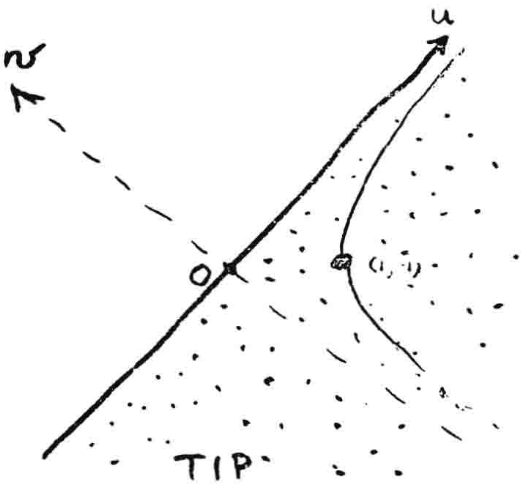

$\notin M$
∞
γ
p
x

$x \in I^-[\gamma]$

future singular point : singular TIP


$\notin M$
γ
x

$x \in I^-[\gamma]$



$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2$$

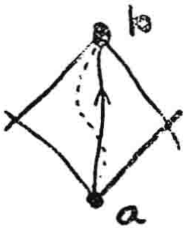
$$u = t - z, \quad v = t + z$$

$$ds^2 = du\,dv - dx^2 - dy^2$$

$$u = -\frac{1}{v}$$

$$\int ds = \int \sqrt{du\,dv}$$

$$= \int_1^{\infty} \frac{dv}{v} = \infty$$

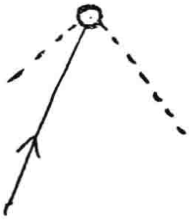


$$\int_a^b ds = \int_a^b \sqrt{dt^2 - g_{ij}\,dx^i dx^j}$$

$i, j = 1, 2, 3$

$$= \int_a^b dt \sqrt{1 - g_{ij} v^i v^j} \leq b - a$$

$v^i = \frac{dx^i}{dt}$



$$ds^2 = dt^2 - dx^2 - dy^2 - dz^2$$

ⓐ regular

$$= \frac{dt^2 - dx^2 - \dots}{\sqrt{t^2 - x^2 - \dots}}$$

ⓑ singular

$$= \frac{dt^2 - dx^2 - \dots}{t^2 - x^2 - \dots}$$

ⓒ at ∞

also: "conical" singularities

New Singularity Theorem:

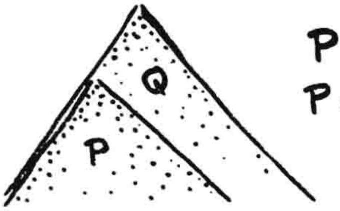
A space-time which

(i)....
(ii)....
(iii)....
and (iv)... (a)...
(b)...
(c)...

} as before (H&P 1970)

must contain a singular TIP or a singular TIF.

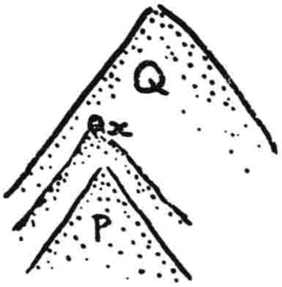
Causality relations on $\hat{M}$ and on $\check{M}$

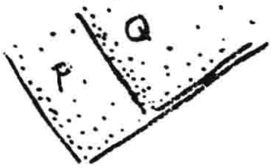


$P \prec Q$ means $P \subset Q$

$P \ll Q$ means $\begin{cases} P \subset I^-(x) \\ \text{some } x \in Q \end{cases}$

$\hat{M}$

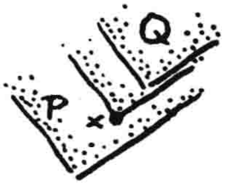




$P \prec Q$ means $Q \subset P$

$P \ll Q$ means $\begin{cases} Q \subset I^+(x) \\ \text{some } x \in P \end{cases}$

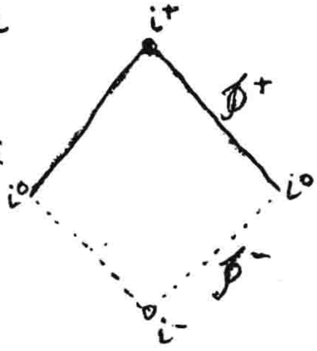
$\check{M}$



Examples

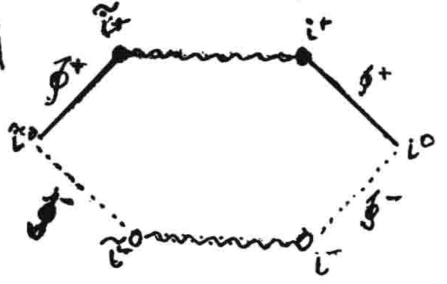
Minkowski

$\hat{M}$


i^+
$\mathscr{I}^+$
i^0
i^0
$\mathscr{I}^-$
i^-

Schwaraschild (Synge-Kruskal)

$\hat{M}$

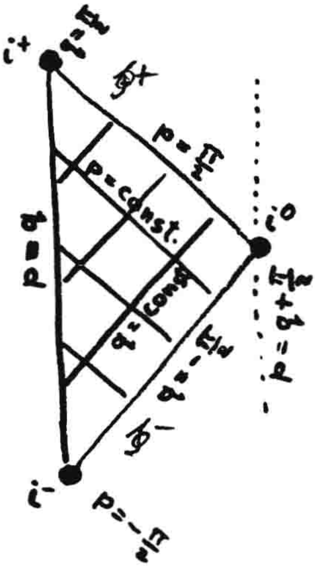

$\tilde{i}^+$
i^+
$\tilde{\mathscr{I}}^+$
$\mathscr{I}^+$
$\tilde{i}^0$
i^0
$\tilde{\mathscr{I}}^-$
$\mathscr{I}^-$
$\tilde{i}^-$
i^-

Conformal Infinity for Minkowski Space-time.

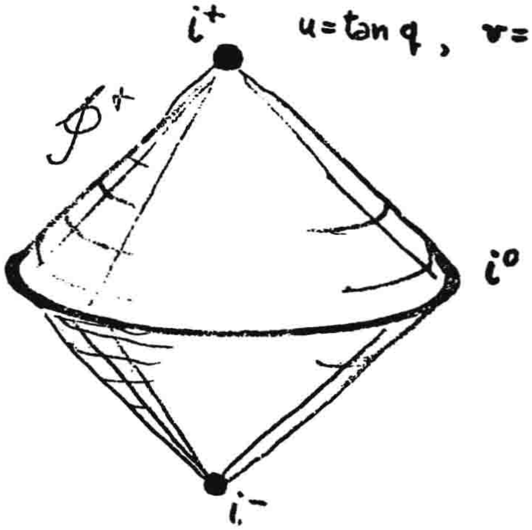
$s^2 = 2\,du\,dv - \frac{1}{2}(u-v)^2(d\theta^2 + \sin^2\theta\, d\varphi^2)$; $u = \frac{t-r}{\sqrt{2}}$, $v = \frac{t+r}{\sqrt{2}}$

$\hat{s}^2 = \Omega^2 ds^2 = 2\,dp\,dq - \frac{1}{2}\sin^2(p-q)\{d\theta^2 + \sin^2\theta\, d\varphi^2\}$; $\Omega = \frac{1}{\sqrt{(1+u^2)(1+v^2)}}$

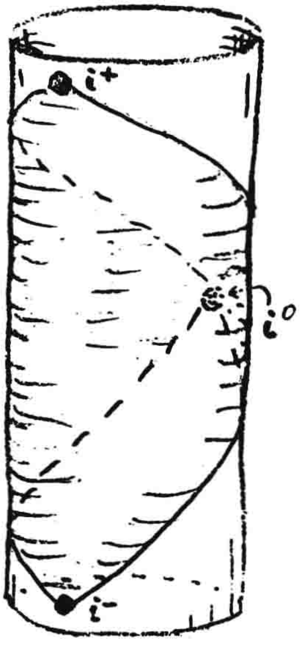
$u = \tan q$, $v = \tan p$

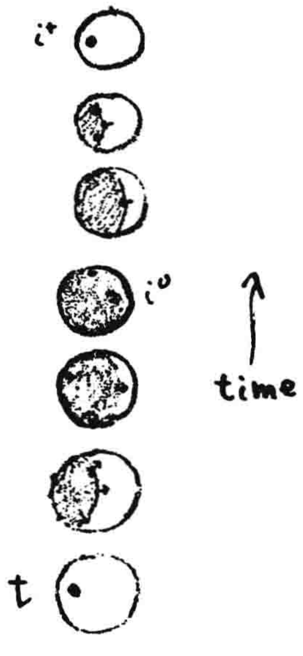


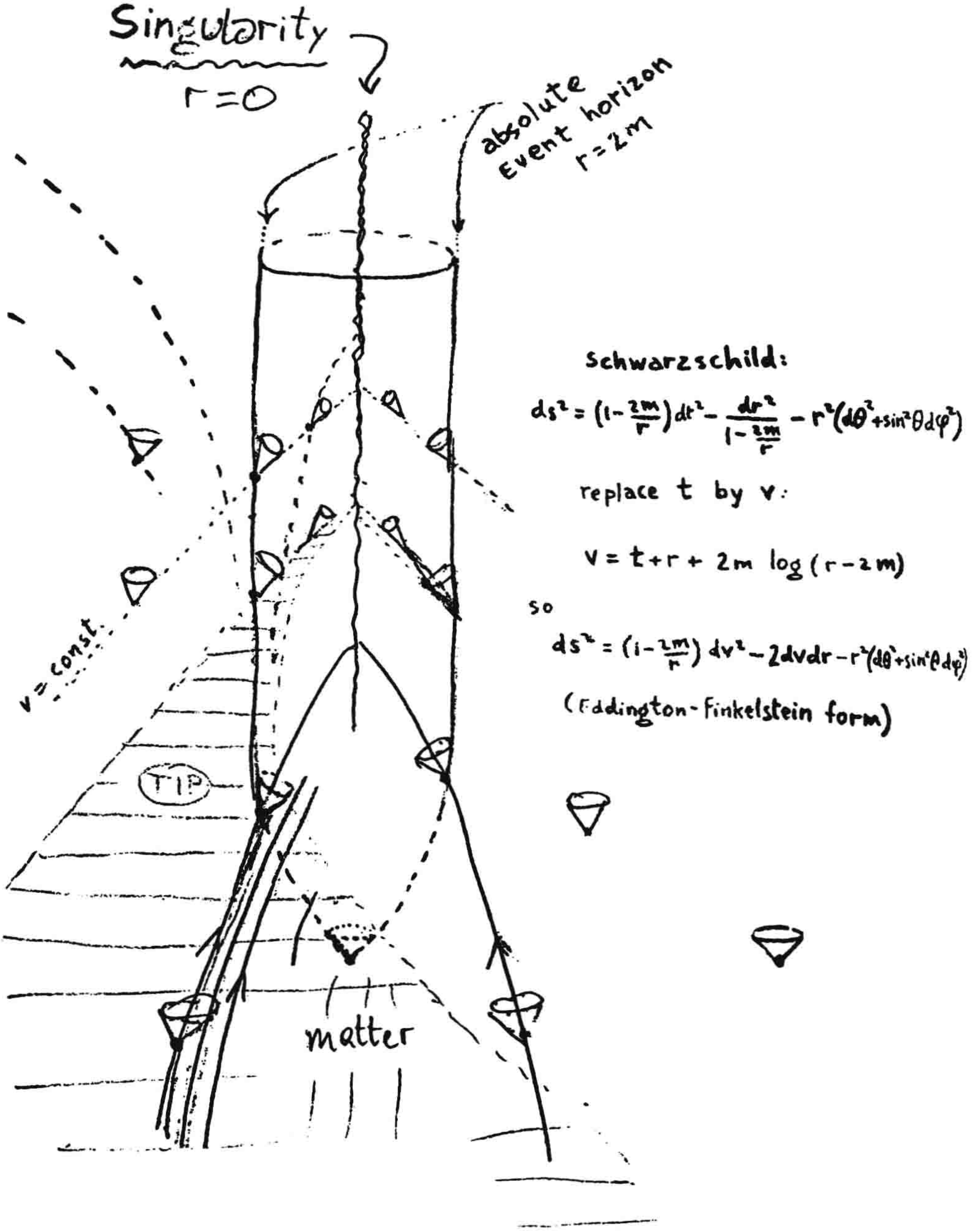
i^+
$q=\frac{\pi}{2}$
$\mathscr{I}^+$
$p=\frac{\pi}{2}$
p=const.
i^0
$p=q$
q=const
$q=-\frac{\pi}{2}$
$p=q+\frac{\pi}{2}$
$\mathscr{I}^-$
i^-
$p=-\frac{\pi}{2}$



i^+
$\mathscr{I}^+$
i^0
i^-






Singularity
r=0
absolute Event horizon
r=2m
v=const.
TIP
matter

Schwarzschild:

$$ds^2 = \left(1-\frac{2m}{r}\right)dt^2 - \frac{dr^2}{1-\frac{2m}{r}} - r^2(d\theta^2+\sin^2\theta\, d\varphi^2)$$

replace t by v:

$$v = t + r + 2m \log(r-2m)$$

so

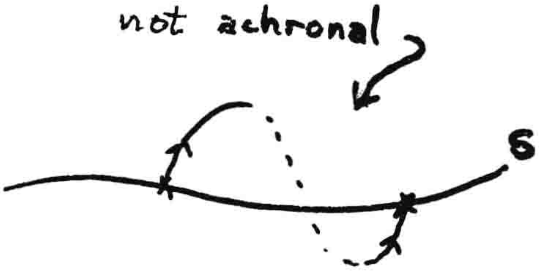
$$ds^2 = \left(1-\frac{2m}{r}\right)dv^2 - 2dvdr - r^2(d\theta^2+\sin^2\theta\, d\varphi^2)$$

(Eddington-Finkelstein form)

S = achronal boundary

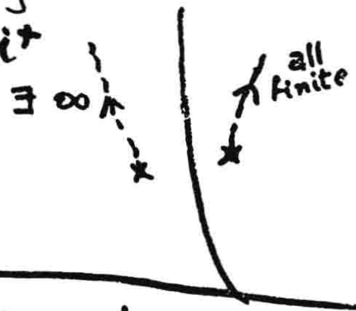
$$S = \partial I^+[Q]$$
$$= \partial I^-[R]$$

S is a topological 3-manifold.



Absolute Event Horizon

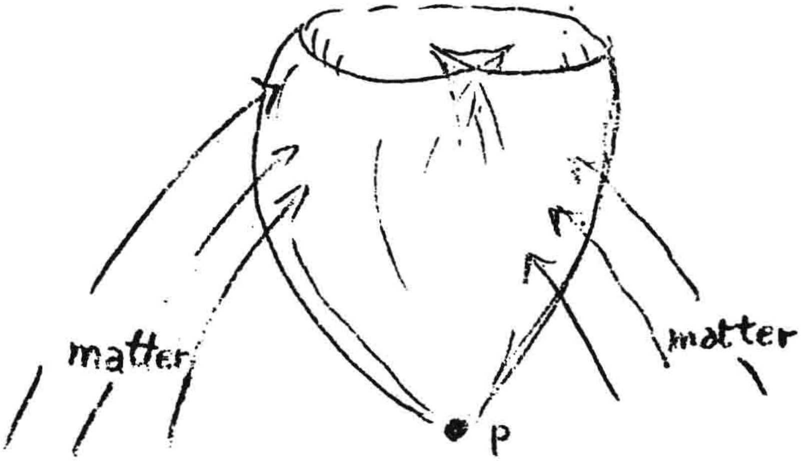
$$= \partial \bigcup \{\text{all } \infty \text{ TIPs at } \mathscr{I}^+ \text{ or } i^+\}$$

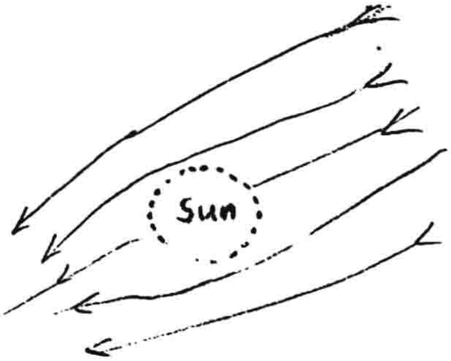


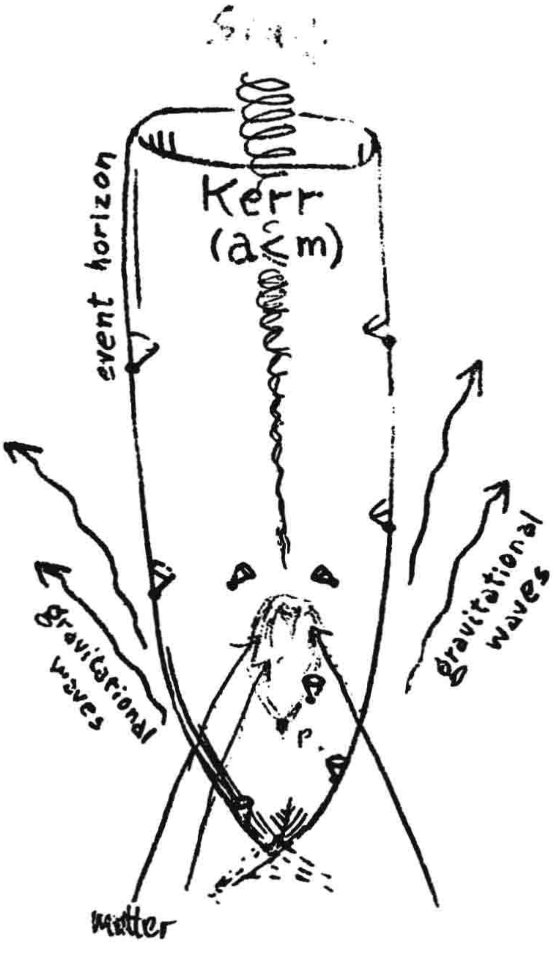
Lemma An achronal boundary is generated by future-endless null geodesics

⟺ it is the boundary of a union of TIPs

Asymmetrical collapse


matter
matter
p




Sing.
event horizon
Kerr
(a≤m)
gravitational waves
gravitational waves
p
matter

$$D^2 A^{\frac{1}{2}} = -\{\Phi + |\sigma|^2\} A^{\frac{1}{2}}$$

$$\Phi = -\frac{1}{2} R_{ab} l^a l^b$$

$$= 4\pi G T_{ab} l^a l^b \geq$$

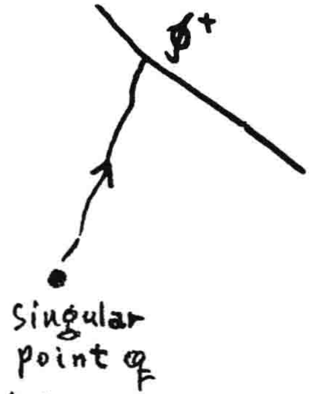
A = element of surfac
area

l^a = null tangent
vector

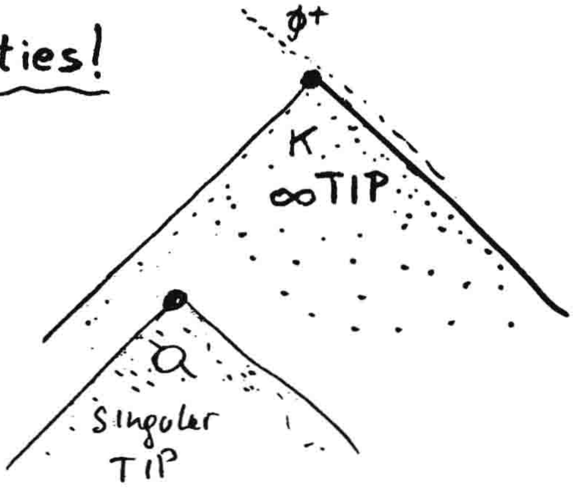
σ = complex sheer

Cosmic Censorship

No naked singularities!

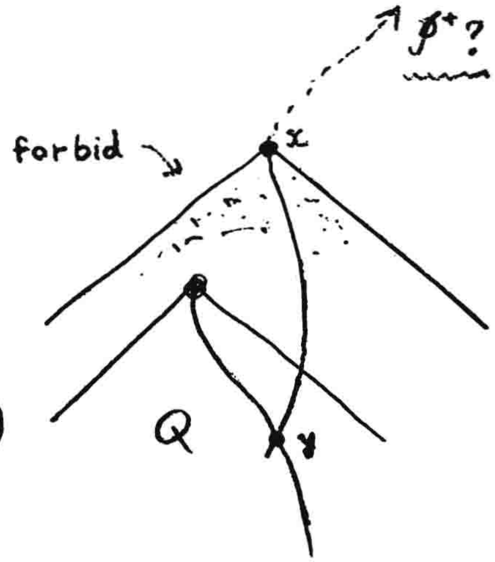

$\mathscr{I}^+$
Singular point q_F

* big bang OK!


$\mathscr{I}^+$
K
∞TIP
Q
Singular TIP

$Q \prec K$

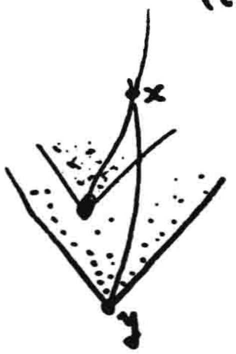
Strong cosmic censorship:


$\mathscr{I}^+$?
forbid
x
Q
y

$\boxed{Q \subset I^-(x)}$

$y \ll q \ll x$

(timelike singularity)

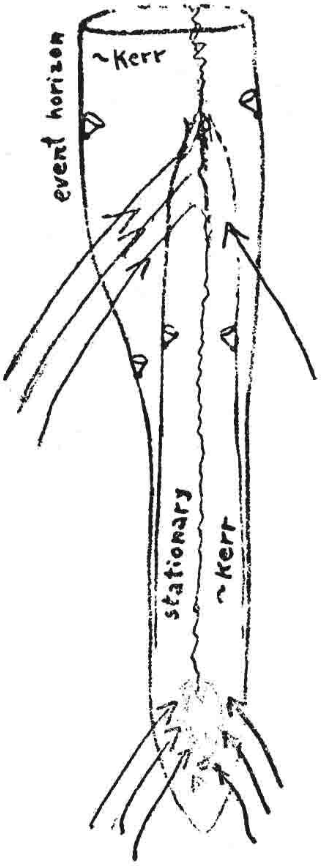


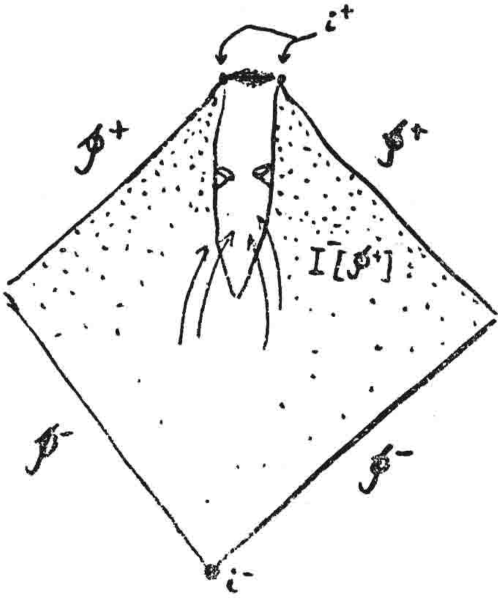
Theorem Strong cosmic censorship

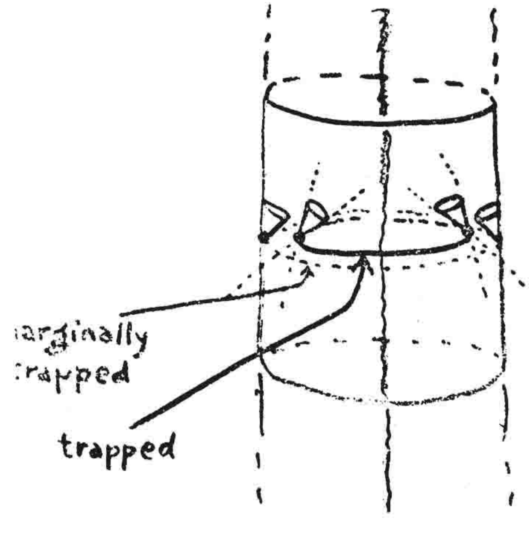
$\Leftrightarrow$ M globally hyperbolic

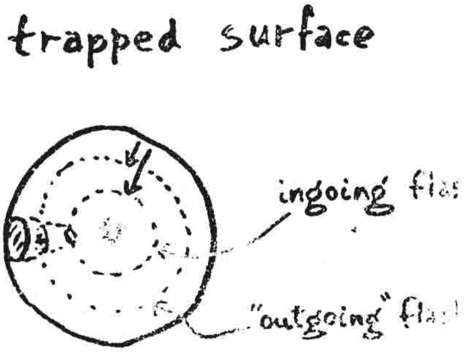
$\underset{\text{(Geroch)}}{\Longleftrightarrow}$ $\exists$ Cauchy hypersurface

(whence topology of universe unchanging for all time)


event horizon
~Kerr
stationary
~Kerr


i^+
$\mathscr{J}^+$
$\mathscr{J}^+$
$I[\mathscr{J}^+]$
$\mathscr{J}^-$
$\mathscr{J}^-$
i^-


arginally
rapped
trapped


trapped surface
ingoing fla
"outgoing" fla

Domains of Dependence & Cauchy Horizons

Take S achronal

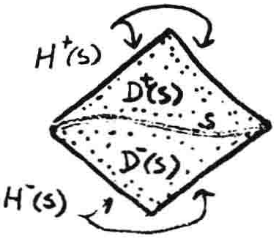
$D^+(S) = \{x \mid$ every past-endless timelike curve through x meets $S\}$

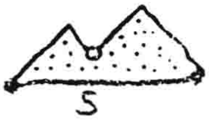
$D^-(S) =$ ----- future ......

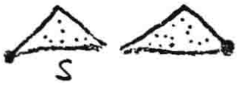
$D(S) = D^+(S) \cup D^-(S)$

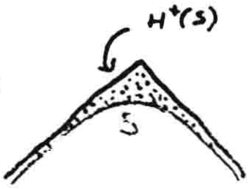
$= \{x \mid$ every endless timelike curve through x meets $S\}$

Hawking takes "causal"





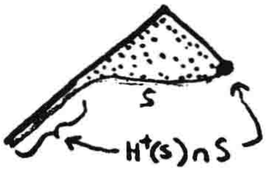


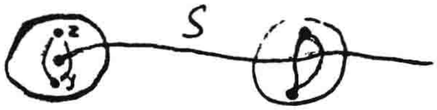


$$H^+(S) = D^+(S) - I^-[D^+(S)]$$

$$H^-(S) = D^-(S) - I^+[D^-(S)]$$

$$H(S) = H^+(S) \cup H^-(S)$$

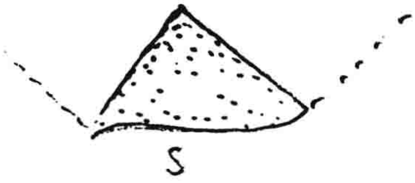




edge $S = \{x \mid$ every neighbourhood of x contains points ~~x~~ z and y and two timelike curves from z to y just one of which meet $S\}$

= (set of limit points of S not in S) ∪ (set of points in whose vicinity S fails to be a a 3 dim. manifold)

Alternative defn of $D^+(S)$



achronal

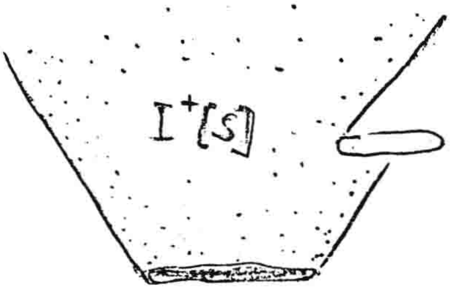
$D^+(S)$ = set of all points of $S \cup I^+[S]$ which do not lie in any TIF of that S

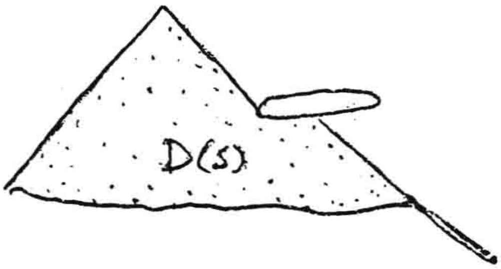
Theorem If $S \subset M$ and $B = \partial I^+[S]$ and $x \in B - \bar{S}$, the

$\exists$ a null geodesic on B with future end-point x and whi

is either past-endless or has past end-point on $\bar{S}$.

Theorem If $S \subset M$ is achronal then every point of $H^+(S) - \text{edge}\, S$ is the future end-point of a null geode on $H^+(S)$ which is either past-endless or has past end-poi on edge S





Hawking's version of Cosmic Censorship

S is a "partial Cauchy surface", i.e. $\boxed{S \text{ is achronal}}$ & $\boxed{\text{edg}\, S = \varnothing}$

(&, normally, S taken spacelike)

M is "weakly asymptotically simple", i.e., for our purposes,

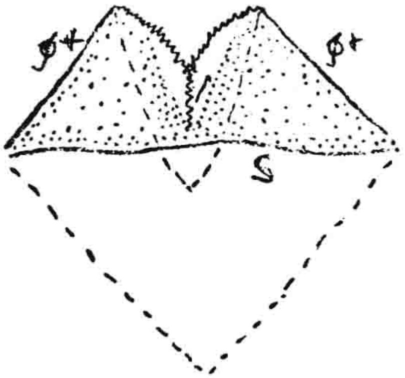
$\boxed{\mathscr{I}^+ \text{ and } \mathscr{I}^- \text{ exist and are as in Minkowski space}}$

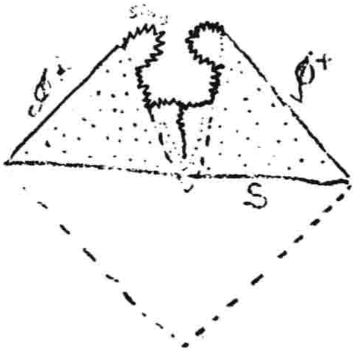
(but $\mathscr{I}^-$ not needed here)

Use causality and chronology of $\hat{M}$ for points of $\mathscr{I}^+$.

M is "(future) asymptotically predictable" if

$\boxed{\text{a partial Cauchy surface } S \text{ exists with } \mathscr{I}^+ \subset D^+(S)}$



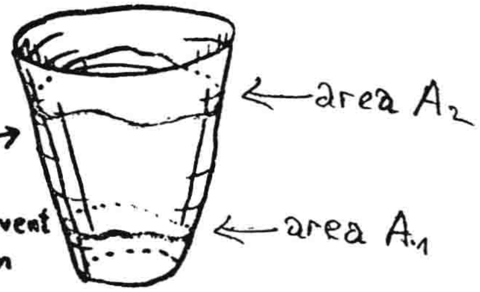


is "strongly (future) asymptotically predictable" if, in addition,

$$\boxed{I^+[S] \cap \partial I^-[\mathscr{I}^+] \subset \text{int}\, D^+(S) \cup \mathscr{I}^+}$$

(Hawking uses $J^\pm$, but makes no difference.)

ʌ use strong Cosmic Censorship by applying



absolute event horizon

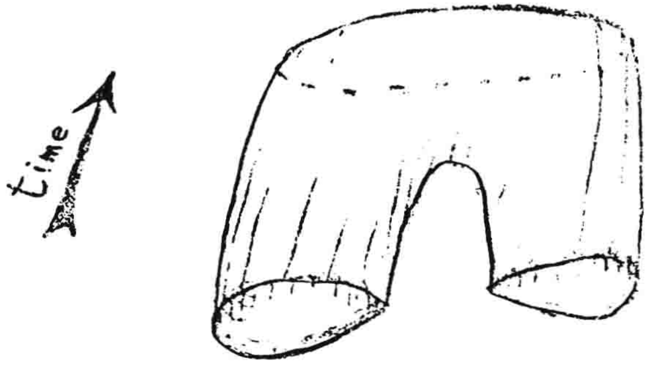
$\partial \bigcup\{\text{all } \infty\text{TIPs}\}$

r $= \partial I^-[\mathscr{I}^+]$

Theorem If the energy condition (i.e. $-R_{ab}t^at^b \leqslant$ for timelike t^a) and Strong C Censorship hold for M, th

$$A_2 \geqslant A_1,$$

where A_2 lies entirely to the future of A_1 along th absolute event horizon.



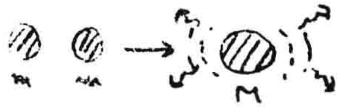
$$\boxed{A \geqslant A_1 + A_2}$$

(Hawking)

$c = G = 1$

Kerr: $A = 8\pi m(m \pm \sqrt{m^2 - a^2})$

(e.g

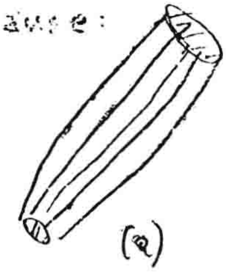


$M < 2m$, $M^2 > 2m^2$ $\therefore \sqrt{2}\,m < M < 2m$

energy radiated $< (2-\sqrt{2})m$.

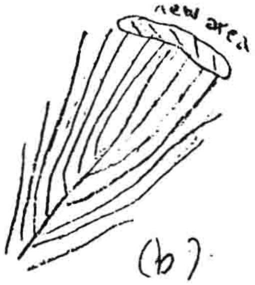
Area increases because:

$$D^2A^{\frac{1}{2}} = -(\Phi + |\sigma|^2)A^{\frac{1}{2}} \leqslant 0$$



(a)

&



(b).

$$dm = \frac{\kappa}{8\pi}\, dA + \Omega\, dJ$$

rtle, Hawking, Carter, Bardeen, Smarr

D. Beckenstein, Phys. Rev. 1973

m = black hole mass
A = " " area
(am =) J = " " angular mom.
κ = " " surface gravity
Ω = " " angular velocity

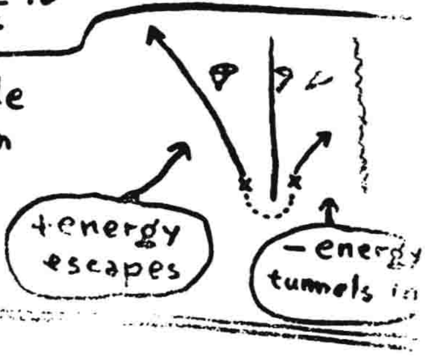
↪ black hole entropy $= (\eta)\dfrac{kc^3}{\hbar G} A$

black hole temperature $= (\eta^{-1})\dfrac{\hbar G}{kc}\kappa$

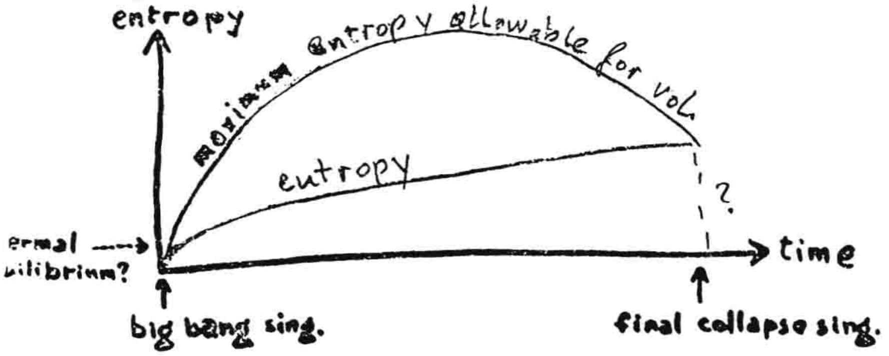
η = num. const. order 1
k = Boltzman's const.

Hawking: tiny black holes explode owing to pair production

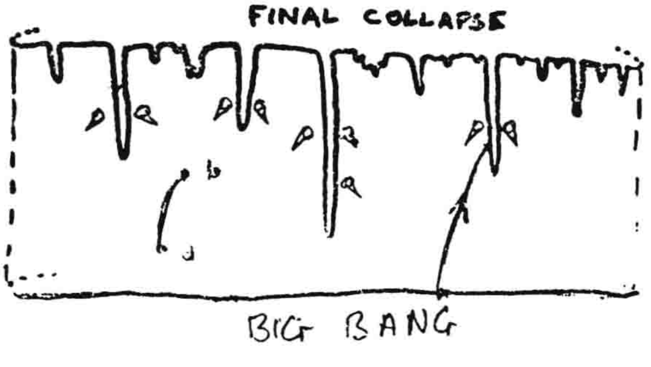
∴ temperature real!



Closed Universe



obally hyperbolic
maximum time
m a to b is attained
varies continuously
th a and b ($a \prec b$).



References

Books

Black Holes
① Les Astres Occlus; Les Houches 1972
(ed. DeWitt & DeWitt; Gordon & Breach 1973)
Espec. article by Hawking & also Carter

② The Large Scale Structure of Space-time, by Hawking & Ellis
(Cambridge Univ. Press 1973)

③ Techniques of Differential Topology in Relativity, by Penrose
(S.I.A.M. Philadelphia 1972)

④ Battelle Rencontres, 1967 Lectures in Mathematics & Physics
(ed. DeWitt & Wheeler; W.A. Benjamin, New York: 1968)

Papers

ⓐ Hawking & Penrose, Proc. Roy. Soc. (Lond.) A314 (1970) 529

ⓑ Hawking, Comm. Math. Phys. 25 (1972) 152

ⓒ Geroch, Kronheimer & Penrose, Proc. Roy. Soc. (Lond.) A327 (1972) 51

ⓓ Penrose, Rivista del Nuovo Cim. Ser. 1 1 Num. Spec. (1969) 252

63
GRAVITATIONAL COLLAPSE

Gravitational Radiation and Gravitational Collapse.
Ed. C. DeWitt-Morette, pp. 82–91. Springer, 1974

This chapter, and also Chapter 65, was a lecture delivered in Poland at a meeting in honour of the 500th anniversary of Nicolaus Copernicus' birth, organized by the International Astronomical Union (IAU).

GRAVITATIONAL COLLAPSE

R. PENROSE

University of Oxford, England

Abstract. In the standard picture of gravitational collapse to a black hole, a key role is played by the hypothesis of cosmic censorship – according to which no naked space-time singularities can result from any collapse. A precise definition of a naked singularity is given here which leads to a strong 'local' version of the cosmic censorship hypothesis. This is equivalent to the proposition that a Cauchy hypersurface exits for the space-time. The principle that the surface area of a black hole can never decrease with time is presented in a new and simplified form which generalizes the earlier statements. A discussion of the relevance of recent work to the naked singularity problem is also given.

The theoretical picture of gravitational collapse to a black hole is now a familiar one (Penrose, 1969; Hawking and Ellis, 1973), so I shall dwell only briefly upon it. The essentials are depicted in Figure 1. Matter collapses inwards under the influence of gravitation until a situation is reached from which there is no escape. We may rec-

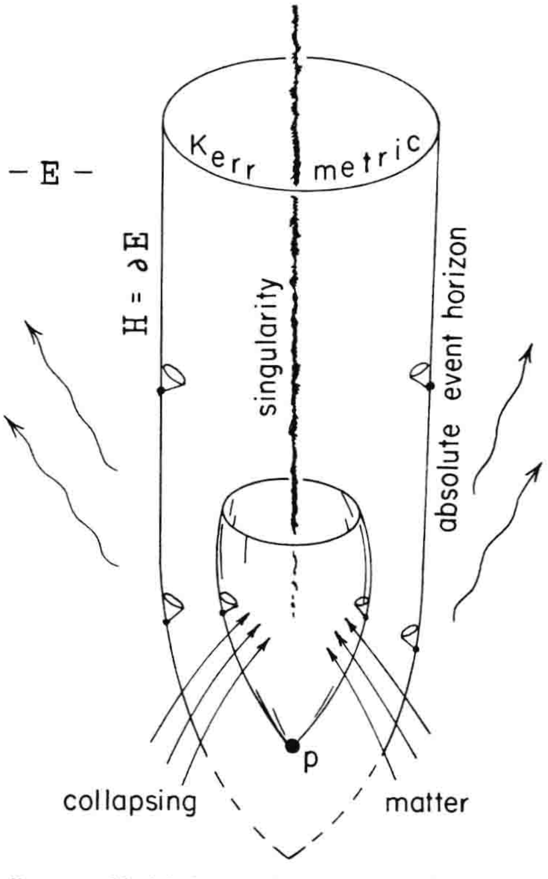


Fig. 1. Gravitational collapse to a black hole according to a standard viewpoint: the future light cone of p starts to reconverge and the external field settles down to that of a Kerr black hole.

C. DeWitt-Morette (ed.), Gravitational Radiation and Gravitational Collapse, 82–91. All Rights Reserved.
Copyright © 1974 by the IAU.

ognise that such a situation has occurred by the presence of a *trapped surface* (Penrose, 1968) or, as in the case depicted, by the fact that sufficient matter crosses the future light cone of some event p to cause the divergence of the light rays (null geodesics through p) to change sign, so that the rays begin to converge towards one another again. When this occurs we can invoke the Hawking-Penrose theorem (Hawking and Penrose, 1970) to deduce the existence of some form of *space-time singularity*. To derive this conclusion we need only assume, in addition to Einstein's equations, that the density plus each principal pressure is non-negative and the density plus the sum of the principal pressures is likewise non-negative, that there are no closed timelike curves and that the space-time is appropriately 'general' – all of which are very reasonable requirements from the physical point of view.

At this juncture the assumption of *cosmic censorship* can be brought in (Penrose, 1969, 1974a). This hypothesis forbids the existence of a *naked singularity*, i.e. any space-time singularity arising in a collapse from which a signal can originate which escapes to infinity. I shall discuss the status of this assumption a little later. For the moment I merely point out that it is the most *conservative* of the positions available to us. The violation of cosmic censorship would lead to a picture of gravitational collapse much more radical than that of a black hole.

Consider, now, the set E consisting of all events from which a timelike curve (or a null curve – it makes little difference) can be drawn into the future to infinity. The boundary ∂E of E is called the *absolute event horizon* (Penrose, 1969). Thus the absolute event horizon may be thought of as the boundary between the regions consisting of those events from which an observer can escape to infinity and those events from which he cannot. The principle of cosmic censorship may now be adopted in the form which states that all the resulting space-time singularities must be surrounded by – and must not lie on – the absolute event horizon $H = \partial E$. One precise statement of this hypothesis is Hawking's condition of future asymptotic predictability (Hawking, 1972; Hawking and Ellis, 1973). This applies, strictly speaking, only in the case of an asymptotically flat space-time. I shall give a more general criterion shortly. When such a non-singular absolute event horizon exists we say that the universe model contains a black hole or black holes – depending upon the number of disconnected pieces into which this horizon falls. We may also envisage two or more black holes congealing into one. Then the absolute event horizon resembles the situation depicted in Figure 2. When a black hole settles down into a stationary state then, as is virtually established by the results of Israel (1967), Carter (1971), Hawking (1972), and others (Müller zum Hagen *et al.*, 1972; cf. also Hawking and Ellis, 1973), it takes up the configuration of a *Kerr* metric. This is defined by just two parameters m and a, where m specifies the mass and am the angular momentum. For a black hole, $m \geqslant a$.

An important feature of black holes subject to cosmic censorship is the area principle. This states that if S_1 and S_2 are two cross-sections of H, where S_2 lies entirely to the future of S_1 along H, then the area of S_1 cannot exceed that of S_2. (Each S_i is a spacelike 2-surface, not necessarily smooth). Hawking (1972) has used the area

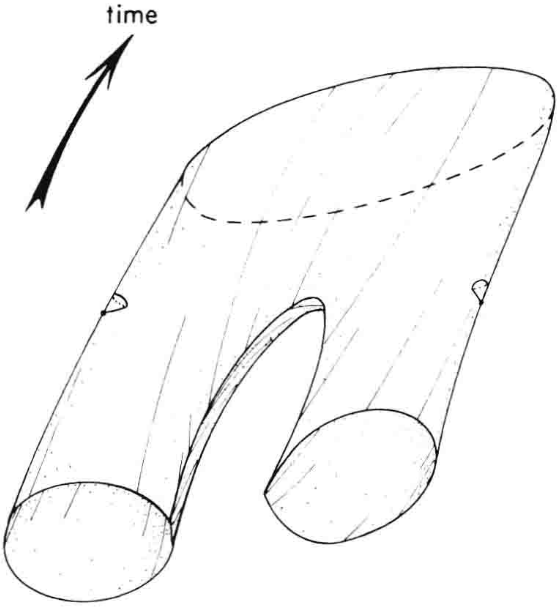


Fig. 2. Two black holes congealing to form a single black hole.

principle to derive an uuper bound to the amount of gravitational radiation emitted when two black holes coalesce. For this, the area principle is used in the form

$$A_1 + A_2 \leqslant A_3,$$

where the subscripts 1 and 2 refer to the colliding black holes and 3 to the resultant black hole, the surface area A_i of each black hole being given by

$$A_i = 8\pi m_i (m_i + (m_i^2 - a_i^2)^{1/2}),$$

when the black holes are settled into their respective Kerr configurations. Thus, in this case the cross-section initially consists of two pieces and finally of one piece, but the area principle still holds.

It is worth indicating here the basic properties which lie behind the area principle. The key fact is that every point of H is the past end-point of some future-endless null geodesic lying on H, of infinite affine extent. If δA is an element of surface area of some spacelike cross-section of H and if we propagate δA along the null geodesic generators of H, we get the equation (cf. Penrose, 1968)

$$-\frac{\mathrm{d}^2(\delta A)^{1/2}}{\mathrm{d}v^2} = (\delta A)^{1/2}\{\sigma\bar{\sigma} + \Phi_{00}\} \geqslant 0,$$

where v is an affine parameter along the relevant null geodesic γ, where σ is its complex shear, and Φ_{00} is minus one-half the Ricci tensor component associated with γ. The *weak energy condition* (energy plus each principal pressure non-negative) gives $\Phi_{00} \geqslant 0$. It follows that if $\mathrm{d}(\delta A)^{1/2}/\mathrm{d}v$ ever becomes negative, then $(\delta A)^{1/2}$ must decrease to zero within a finite v value. But if this were to happen, then γ could not remain on H for values of v beyond this (H being the boundary of a set, namely E, which is its own past, cf. Penrose, 1972). This would contradict the statement that each null geodesic generator of H extends into the future to arbitrarily large affine parameter values. Thus $\mathrm{d}(\delta A)^{1/2}/\mathrm{d}v \geqslant 0$, whence $\mathrm{d}(\delta A)/\mathrm{d}v \geqslant 0$ along γ.

This shows that the surface area increases along γ as we proceed into the future. Furthermore γ never leaves H in *future* directions. But it can leave H in *past* directions. Thus, the surface area of cross-sections of H can increase into the future for two reasons, either because it increases along any *given* γ or because *new* γ's emerge. (Note that in the black hole collision of Figure 2, both of these situations occur.) Neither of these situations can occur in past directions. Thus, the area principle is established provided it can be shown that null geodesics generate H, having infinite affine lengths into the future. In fact, it is only the condition of infinite affine length which causes any difficulty. It is this that requires a suitable statement of the cosmic censorship principle.

In order to state this principle precisely we require a definition of a space-time singularity. I shall adopt a definition which is not quite the same as that suggested by the singularity theorems, but it is more useful for present purposes. The ideas are taken from Geroch *et al.* (1972) and are somewhat similar to those of Seifert (1971). Let us assume, for simplicity, that the space-time is strongly causal (cf. Hawking and Ellis, 1973; Penrose, 1972). We adjoin some extra 'ideal points' to the space-time which are end-points to future-endless or to past-endless causal (i.e. timelike or null) curves. We say that two future-endless causal curves γ, γ' have the same ideal future end-point if the two curves have identical pasts (i.e. $I^-[\gamma]=I^-[\gamma']$ in the notation of Geroch *et al.* (1972). Such sets are called terminal indecomposable past sets, or TIP's for short). Similarly, two past-endless causal curves have the same ideal past end-point if they have identical futures (called TIF's). (To be able to say when a future-endless causal curve has the same future end-point as the past end-point of some past-endless causal curve would require some extra complications which I do not propose to enter into here.) These ideal points form a sort of boundary to the space-time. But the boundary points need not be singularities. They may be points at infinity. To distinguish these two possibilities we may adopt one of a number of slightly different criterea. Let us choose the simplest one here and say that the boundary point is at infinity if and only if there is a semi-infinite causal curve of infinite proper length which has that boundary point as an ideal end-point. Otherwise, we can say we have a finite boundary point. In the case that we have a space-time which is maximally extended (i.e. not a proper part of any other connected space-time), then all *finite* boundary points may be reasonably interpreted as *singular* points of the space-time. One might also consider some of the points at infinity also to be singular, but I shall not bother with this possibility here.

Let us now try to interpret the cosmic censorship hypothesis in these terms. We need a definition of a naked singularity, but we must be careful not to rule out the 'big bang' in the exclusion of naked singularities, even though the big bang singularities of the normal cosmological models are indeed naked in the sense that signals can escape from them (i.e. they are represented, in the above descriptions by past end-points of past-endless causal curves). The idea of a naked singularity is that it might arise in the collapse of ordinary matter from a non-singular initial state. I shall formalize this idea by regarding a singularity as naked if there is some observer (timelike curve) for whom the singularity lies initially to his future and subsequently to his past. This concept is in essence time-symmetrical, and it should be observed that the normal cosmological big bang does not qualify as a naked singularity in this sense. The definition is basically 'local' in that no mention is made of signals escaping to infinity. Thus, even a singularity inside a black hole might conceivably be 'naked' to some observer who is himself inside the black hole. However, in the normal picture of spherically symmetrical collapse, such a situation does not occur. There is some indication also (Simpson and Penrose, 1973) that in a generic perturbed collapse this situation still does not occur.

Now we can apply the above idea to the aforementioned definition of a singularity either of two ways around. Let us suppose, first, that we are concerned with the ideal *past* end-point q of a past-endless timelike curve γ (i.e. with the TIF $I^+[\gamma]$). We can say that q lies to the (causal) future of a point p if the future of p contains the future of γ (i.e. $I^+(p) \supset I^+[\gamma]$). In this case, any point r of γ lies to the future of p, so there are timelike curves from p to r. An observer following one of these timelike curves will have q to his future when he is at p and to his past when he is at r. (See Figure 3.)

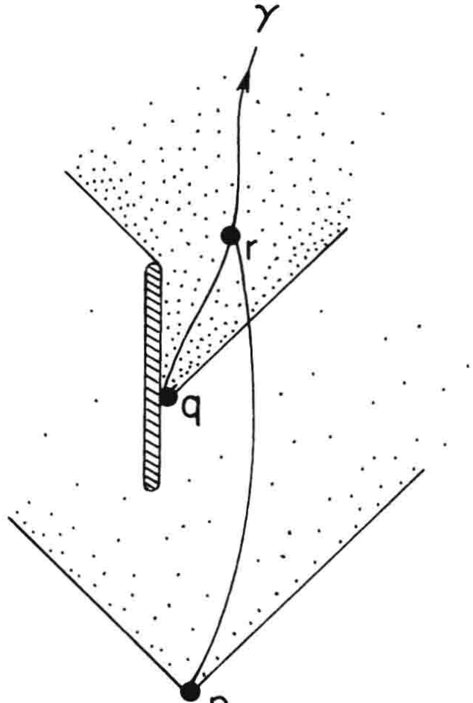


Fig. 3. A naked singularity q – the future of some point p contains the future of q.

Thus, if q is a singular point, then this is the situation where a naked singularity occurs. A similar situation might possibly arise when q is a point at infinity. For purposes of the present discussion I want to regard such a possibility as something akin to a naked singularity. The normal cosmological models do not admit such an occurrence. (The complete anti-de Sitter space is an exception, cf. Penrose (1968), but cannot be regarded as a 'normal' cosmological model.) Also, while this possibility *can* occur in certain idealized models of collapse, there is some indication that it may be unstable in the sense that small perturbations could lead to genuine curvature singularities (cf. Simpson and Penrose, 1973). I shall refer to the hypothesis which excludes the above situation from occurring whether or not q is a point at infinity as *strong cosmic censorship.*

We could also apply the above argument in a time-reversed sense, allowing q to be the future end-point of a future-endless timelike curve γ'. It turns out (although this is not immediately obvious) that we get precisely the same condition of strong cosmic censorship as above if we proceed in this time-reversed way, i.e. if we rule out any point r whose past could contain the past of such a curve γ'. In fact, it can be shown (cf. Penrose, 1972) that strong cosmic censorship is equivalent to Leray's condition of *global hyperbolicity*. Using a theorem of Geroch (1970) we can state global hyperbolicity to be the condition that a Cauchy hypersurface exists for the space-time. Geroch's theorem also implies that the entire universe model is then topologically the product of this Cauchy hypersurface with an open timelike line. Thus, we may say that if the strong cosmic censorship principle holds, then the topology of the universe is unchanging for all time. (I shall discuss this more fully, in the case of a closed universe model (Penrose, 1974b)). In fact, global hyperbolicity leads to many other simplifications in the global structure of a space-time. For example, any two points with a timelike separation in such a space-time can be connected by a timelike geodesic which maximizes the lengths of all timelike curves which connect the points. Thus, if strong cosmic censorship turns out to be a reasonable requirement for realistic space-time models, then many global problems connected with cosmology and black-hole structure can be handled without fear of the sort of pathology arising which could complicate so much of the earlier global analysis.

As one example of the utility of strong cosmic censorship, let me give a statement of the area principle which generalizes the particular version for asymptotically flat space-times given by Hawking (1972) (cf. also Hawking and Ellis, 1973). Suppose the space-time is globally hyperbolic and the energy condition holds everywhere. Define E to be the set of points which are past end-points of timelike curves of infinite length. The required result is that the *absolute event horizon* ∂E then satisfies the area principle (i.e. if S_1 and S_2 are two sections of ∂E, with S_2 to the future of S_1, then the area of S_2 is not less than that of S_1). To specialize this result to the situation considered by Hawking, namely to a space-time which is future asymptotically predictable from some spacelike hypersurface K, we need consider, for our space-time manifold, only that portion of the entire space-time which is the interior of the domain of dependence of K (cf. Hawking and Ellis, 1973; Penrose, 1972). But the generalized version of the result as stated here may be applied also to cosmological situations. I shall give a proof and discussion of the result elsewhere.

What is the theoretical evidence in favour of or against cosmic censorship (not necessarily in its strong form)? In my opinion the evidence in either direction is very scanty indeed. I think that in a collapse which does not differ very greatly in its initial conditions from the standard spherically symmetrical collapse situation I described first, then some form of cosmic censorship is likely to be valid. The analysis by various workers of the stability of the Kerr metric tends to confirm this belief. (Much of this work is being described elsewhere in this symposium (cf. Chandrasekhar, 1974; Press, 1974; Teukolsky, 1974).) It seems likely that this programme will come to a successful conclusion in the near future and that the Kerr solution (for all m, a with $m>a$) will be declared 'officially' to be stable. Perhaps I can be forgiven, on the other hand, if I pose a few queries concerning the whole question of stability for a black hole. I am not trying to cast any doubt on the existing analysis but merely suggest some further questions that one might attempt to answer.

In the first place, it is not completely clear to me that it is legitimate to assume, in the context of establishing that all solutions settle into a Kerr configuration, that a black hole is ever exactly stationary. I have in mind the situation which arose in connection with the Newmann-Penrose constants for asymptotically flat space-times. In fact, one can show quite rigorously that a transition from one exactly stationary state to another exactly stationary state via an intermediate phase in which gravitational radiation is emitted cannot occur unless a certain combination of multipole moments returns to its original value (Newman and Penrose, 1968). But this does *not* mean that this multipole moment combination must return to its original value in order for the system to settle down. It is simply that the presence of a small and ever-decreasing disturbance in the distant gravitational field (the backscattered gravitational radiation) is always just sufficient to prevent any inference concerning the final multipole moment combination to be drawn. I think it is unlikely that anything of this nature could arise to spoil the black hole results, but the issue is not clear to my mind.

It is also not quite clear to me whether the perturbation analysis is actually aimed at excluding the possibility that the absolute event horizon might itself develop into a curvature singularity – which would then be a naked singularity. A 'perturbation' which is actually singular on the event horizon might be excluded on the basis that it represents unreasonable initial data with which to start, but this becomes less clear if the black hole is never assumed to have quite reached a stationary state.

These points are perhaps subtleties and may well have no real significance for the question of black hole stability. However, a more serious question concerns the stability of a black hole near the critical case $a=m$. If a is only marginally smaller than m and a perturbation is applied which is in some sense comparable with the difference $m-a$, then it is by no means clear that such a perturbation can be regarded as small. It is difficult to see how this sort of disturbance could be analyzed within the framework of perturbation theory.

Of course, the whole question of large disturbances applied to black holes is quite an open one. For example, it is often assumed that if two black holes of comparable mass are brought together, then the result will again be a black hole. While I would agree that this is certainly one clear possibility, it is by no means obvious that it is the only one. The possibility that two black holes might collide to form a naked singularity is excluded *only* by the pure assumption that the cosmic censorship hypothesis holds. This would again be a question of an absolute event horizon developing into a naked singularity – a circumstance which I do not see how to exclude on theoretical grounds. Perhaps the computer analysis by DeWitt and his co-workers of a black hole head-on collision will shed some light on this question. It is also possible that two black holes spiralling into one another might produce a qualitatively different result from a head-on collision.

In view of the above uncertainties it is worthwhile to investigate whether or not it is possible to set up initial states of collapse for which a contradiction with the standard picture might be obtained. The combination of area principle, mass-energy conservation, and the final situation of a Kerr metric, together impose definite constraints on the initial geometry. Basically, one is not allowed an initial situation involving a trapped surface of too great area for the initial mass involved. I have described elsewhere (Penrose, 1974a) an attempt at obtaining a contradiction with the standard picture (and hence with cosmic censorship) in this way. Some partial results by Gibbons (1973) have made it seem unlikely that a contradiction with the standard picture can be arrived at by such considerations alone.

A more direct attack on the naked singularity question is that of Müller zum Hagen *et al.* (1973). They construct an explicit solution of Einstein's equations which describes a collapsing dust cloud which eventually forms a black hole. However, before doing so, the dust encounters caustics at which the density becomes infinite. For a short while this region of infinite density – and therefore infinite curvature – is visible from infinity (Figure 4). Thus the solution must be said genuinely to describe a collapse with a naked singularity. It seems that the solution should be stable under small perturbations of the initial state since Grischuk (1967) has shown that fully general timelike singularities (involving rotation) can occur with dust. Müller zum Hagen *et al.* (1973) also show that in the spherically symmetrical case, the introduction of a bounded pressure does not substantially affect their conclusions: a naked singularity can still arise. Stability of the singularity situation for such a fluid under general perturbations (which involve some rotation) is not considered however.

I think that these examples are interesting more for the questions that they raise than for the questions that they answer. I certainly do not feel that in themselves they overthrow the cosmic censorships principle, but they do cause one to wonder what form of precise statement such a principle should have, if one is to have any hope of proving it. Naked singularities which are not stable under perturbations of the initial conditions should presumably be discounted. But what about perturbations of the equations of state? What kinds of such perturbation should be permitted? Or would it be simpler to restrict attention just to the vacuum case? What role should be played by considerations such as those of Hagedorn (1968) according to which the maximum permissible pressure-to-density ratio goes down to zero as the density increases to infinity? If this is valid, the material would behave more and more like dust as the density increases.

Finally, we should ask what is the observational status of naked singularities.

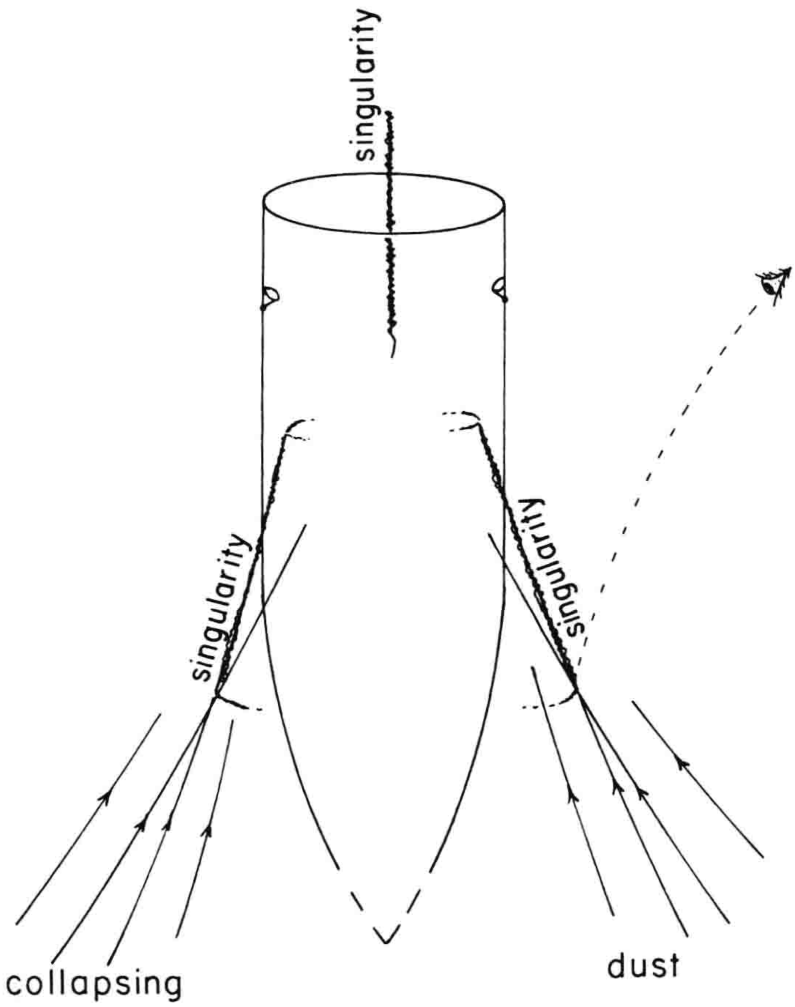


Fig. 4. The Müller zum Hagen-Seifert-Yodzis model of a collapsing dust cloud with naked singularity of the 'shell-crossing' type.

There are certainly many strange phenomena observed to occur, especially in galactic nuclei and in quasars. The temptation to invoke some sort of naked singularity as an explanation must be strong. But the reasons for doing so would seem to be of a rather negative character. The physics at a naked singularity would be largely unknown, so explanations of this kind would have little predictive power. I feel that there is one possible exception, however, and this is in Weber's observations. If his results turn out to be substantiated, then we shall be faced with a problem of energy balance which seems to have little hope of solution unless a very effective beaming mechanism can be suggested which sends the gravitational waves out closely in the plane of the galaxy. Misner's ingenious suggestion (cf. Misner, 1974) for doing this by means of gravitational synchrotron radiation from particles orbiting a black hole seems now not to be feasible. There is an alternative possibility whereby a naked Kerr ring singularity achieves such beaming (Penrose, 1974a) and this should also be considered. But only if the observations pointed clearly to the necessity of such an explanation could one be expected to take such suggestions very seriously.

References

Carter, B.: 1971, *Phys. Rev. Letters* **26**, 331.
Chandrasekhar, S.: 1974, this volume, p. 63.
Geroch, R.: 1970, *J. Math. Phys.* **11**, 437.
Geroch, R., Kronheimer, E. H., and Penrose, R.: 1972, *Proc. Roy. Soc. London* **A327**, 545.
Gibbons, G.: 1973, to appear.
Grischuk, L. P.: 1967, *Sov. Phys. J.E.T.P.* **24**, 320.
Hagerdorn, R.: 1968, *Nuovo Cimento* **56A**, 1027.
Hawking, S. W.: 1972, *Comm. Math. Phys.* **25**, 152.
Hawking, S. W. and Ellis, G. F. R.: 1973, *The Large Scale Structure of Space-Time*, Cambridge Univ. Press.
Hawking, S. W. and Penrose, R.: 1970, *Proc. Roy. Soc. London* **A314**, 529.
Israel, W.: 1967, *Phys. Rev.* **164**, 1776.
Misner, C. W.: 1974, this volume, p. 3.
Müller zum Hagen, H., Robinson, D. C., and Seifert, H.-J.: 1972, *Gen. Relativity and Gravitation* **4**, 1, 53.
Müller zum Hagen, H., Seifert, H.-J., and Yodzis, P.: 1973, *Comm. Math. Phys.* **34**, 135.
Newman, E. T. and Penrose, R.: 1968, *Proc. Roy. Soc. London* **A305**, 175.
Penrose, R.: 1968, in C. M. De Witt and J. A. Wheeler (eds.), *Battelle-Rencontres*, Benjamin, New York.
Penrose, R.: 1969, *Rivista del Nuovo Cimento, Ser. 1*, **1**, Num. Spec. 252.
Penrose, R.: 1972, *Techniques of Differential Topology in Relativity*, S.I.A.M., Philadelphia.
Penrose, R.: 1974a, *Ann., N.Y. Acad. Sci.* **224**, 125.
Penrose, R.: 1974b, in M. S. Longair (ed.), 'Confrontation of Cosmological Theories with Observational Data', *IAU Symp.* **63**, in press.
Press, W. H.: 1974, this volume, p. 93.
Seifert, H.-J.: 1971, *Gen. Relativity Gravitation* **1**, 247.
Simpson, M. and Penrose, R.: 1973, *Int. J. Theor. Phys.* **7**, 183.
Teukolsky, S.: 1974, this volume, p. 92.

64

THE ROLE OF AESTHETICS IN PURE AND APPLIED MATHEMATICAL RESEARCH

Bull. IMA **10** (1974), 266–71

On p. [269] we find the first aperiodic prototile set (Figure 5) which assembles to form the 5-fold quasi-symmetric pattern. This is shown in Figure 4, the pattern being necessarily globally arranged according to the hierarchical scheme indicated in Figure 6 on p. [270]. This was the starting point of the route to the so-called 'Penrose tiles' described in Chapter 111.

The Rôle of Aesthetics in Pure and Applied Mathematical Research

R. PENROSE, FRS, FIMA
Mathematical Institute, Oxford

IT is with great sadness that I have learnt of Professor Coulson's death. In the brief period for which I had been acquainted with him I had come to regard him as a close friend. He was to have talked about the justification of doing research. My title had been chosen so that my talk might have had relevance either to that question or to the question of how research is carried out. I hope now that the relevance of my talk to the former question is something more than superficial—although I cannot in any way hope to fill that vacuum left by Charles Coulson.

I think when we talk about the justification for doing mathematical research, first we have to ask what is the justification in doing mathematics at all. Then we can divide mathematical activity into that which is concerned with existing mathematics and that concerned with as yet non-existing mathematics. Indeed, the two main jobs carried on in mathematics departments at universities are teaching, which has to do with the conveying of the understanding of existing mathematics, and research, which aims at reducing the area of non-existing mathematics. Then, so long as it is true that there is justification in that body of knowledge and understanding which is the mathematics that exists today, then surely there is justification in increasing that body. The dividing line between the known and the unknown that now exists can have no absolute significance. So the real question would be, how does one justify mathematics?

There are two main considerations usually mentioned in this connection, namely the value of mathematics as something which is studied for its own sake and as something studied for the sake of its applications to something else. Of course, there are also the applications of mathematics to itself. Often, in mathematics, one talks about whether a result is useful or not, meaning whether it is useful to other areas of mathematics. To a working mathematician this seems worthwhile in itself and he does not feel that he has to justify it especially. But I think if you talk to the man in the street you would find that the reasons he thinks that mathematics is worthwhile are probably quite different from those that people who actually study mathematics or do mathematical research feel are the reasons. The man in the street might feel that the justification lay solely in being able to do bigger and better calculations, so that stronger and cheaper bridges or aeroplanes etc. could be constructed. Of course this is part of the story, but for a mathematician it is only a rather small part. Indeed, I think that there is not much difference between the motivation involved when mathematics is applied within itself and generally when it is applied in other fields of science.

Basically, the motivations turn out often to be ultimately aesthetic ones; so often a subject is pursued simply for the pleasure that it gives. Perhaps this is a pleasure that one feels because one is searching for truth. Certainly in mathematics there is a supreme requirement of truth which does not exist in other fields, or at least not to the same extent. But truth in itself is by no means sufficient. One would not want to have journals full simply of multiplication tables, say, even although that would be something which might be just as true as (or often more true than) what one normally finds in journals. (At least one could more readily check that the results were true in the case of multiplication tables than in the case of what is normally written in mathematical papers!) So I think it is clear that, although in mathematics truth is a necessary condition, it is far from being a sufficient one. How, in fact, does one decide which things in mathematics are important and which are not? Ultimately, the criteria have to be aesthetic ones. There are other values in mathematics, such as depth, generality and utility. But these are not so much ends in themselves. Their significance would seem to rest on the values of other things to which they relate. The ultimate values seem simply to be aesthetic; that is, artistic values such as one has in music or painting or any other art form.

This, of course, is nothing to be ashamed of. But in mathematics, particularly, there is a difficulty, especially if you are going to ask governments to support research, because the audience is such a very small one. I think that if the subject is to be judged solely on this basis, one would have to compare it, perhaps, with some very off-beat, *avant-garde* form of art where the people who could really understand what was going on were very few indeed. I think if one really needed to justify mathematical research on these grounds it would be extremely hard to make out a good case that there was a large enough group of people who could adequately understand any given area of current mathematical research. But set against the matter of the very small audience, mathematics does have the advantage that directly or indirectly, and often in unexpected ways, it has applications in many other fields. So one has to stress this question of applications. But, although in other subjects the criteria of importance are not quite the same as in mathematics, often they are aesthetic criteria again. For example, I was talking recently to a theoretical physicist and was interested to hear that the thing he would most like somebody to say about one of his papers was that it was elegant. The consideration you might have thought he would regard as most important would be the extent to which his work sheds light on what is going on in the universe. But really these criteria, when it comes down to it, are again very much aesthetic ones. Even in experimental work, aesthetic criteria are very important. After all, you can talk about a beautiful experiment. I think this is certainly something that is strived for. It may not be the only thing, but to have an experiment which for some reason has an elegance about it is, I think, regarded as a great virtue.

One must also appreciate that there are really several rather different kinds of aesthetic criteria even within mathematics itself. For example, the question of whether a theorem is beautiful is not quite the same as the question as to whether the proof of a theorem is beautiful. I feel very often that in mathematics it is the proofs that are held in highest esteem as things of beauty, rather than the theorems themselves. And the beauty in the content of a theorem or the attractiveness of the abstract ideas contained in a proof are different again from elegance in presentation of a result. Again, within any one of these categories there would be different ways of judging. You might regard something as just pretty, for instance, or you might perhaps regard it as (for want of a better word) mind-blowing. An argument which uses logic in an unexpected way can have a special appeal in that way. I am thinking of the sort of discussion involved in connection with the paradoxes that one gets in mathematical logic. There is a beauty in that which is of quite a different character from that which you might find in just an elegant proof of a result, so I think there are many facets to all this.

Basically, I have been talking about pure mathematics. In applied mathematics, I think the question of aesthetics is certainly not so important because there are also other criteria of relevance. One may be interested just in the solution of a problem and be prepared to use any method that one can lay hands on in order to obtain the required result. The question of polishing up the arguments after the result has been obtained may be of no concern. But it is certainly true that there are aesthetic criteria in applied mathematics too. Often, one is not concerned *only* with getting results, but is interested also in putting the thing in an elegant form. Indeed there is also a danger of a tendency to modify the problem in order that a nice argument will apply although when I say it is a danger I do not mean to imply that it is always a danger; it is sometimes a good thing. It is very hard to sort these questions out. But often I think there *is* a danger of just finding elegant solutions to artificial problems; the problems just have to fit in with the solution rather than the solution being suitable for the problem. This is often because the problem is much too difficult to handle by existing techniques. It is possible that with computers the bias will shift, so that one may be less interested in finding nice closed form solutions for the answer to a problem. Having such things now as visual displays and other ways of presenting data I think the aesthetic criteria would again apply but in a different way—perhaps in a more relevant way too. So this is something you will probably see more of in the future when computers get better and the techniques get better for putting problems on the machines and for getting the results from the machines. Perhaps one's aesthetic judgements will change as well. Such judgements are, in any case, often to a considerable extent, acquired tastes. In these cases one cannot really appreciate the beauty of something until some familiarity with it has been gained—one has really to have thought about it for quite a while. This could apply here too. I think that the criteria for deciding what constitutes a nice solution to a given applied mathematical problem may perhaps shift from what it used to be in the more classical approach to applied mathematics. This is *one* aspect of applied mathematics. There are several other aspects, some of which perhaps should be classified more as theoretical physics. It is really more that sort of area which has been my own professional interest. I have not myself been so much concerned with the more classical applied mathematical type of problem and I would like to say more, shortly, about the theoretical physics aspects.

So far I have been considering questions of aesthetics as an end. That is, I have been concerned with aesthetics in its rôle as a justification for what one is doing. There is also a more subtle rôle played by aesthetics, in connection with research, namely as a means for obtaining results. This is really an important aspect of the rôle of aesthetics which I do not think has been discussed very much. It is a mysterious thing in fact how something which looks attractive may have a better chance of being true than something which looks ugly. Of course there is more to it than that, and one can be easily misled, but there seems to be something profound in this and it would be nice to understand more about it. It is the kind of thing that is hard to test in any objective way, but I am sure that successful research workers in practically any field would be familiar with this phenomenon. I have noticed on many occasions in my own work where there might, for example, be two guesses that could be made as to the solution of a problem and in the first case I would think how nice it would be if it were true; whereas in the second case I would not care very much about the result even if it were true. So often, in fact, it turns out that the more attractive possibility is the true one—or that while thinking about the problem in this way the true solution would finally emerge to reveal itself as even more attractive than either contemplated earlier. This is not always the case, of course, but it seems to be much more than just chance. Most guesses are likely to be wrong anyway. But the proportion of correct guesses among the attractive possibilities seems to be far higher than among the less attractive ones.

This is also very true in physics. There was a statement made to me by a colleague of Dirac's who had apparently asked Dirac how it was that at a time when all the great physicists were trying to see how to combine relativity and quantum theory and were all failing, Dirac was able to produce his equation. Dirac had replied "I have a very keen sense of the beautiful and when I found my equation I knew I was right." The question is: how did he know he was right just by the look of the equation? Well I am sure it was not just the *look* of the equation. There must have been many other considerations that went into this, but they all went together to make up an aesthetic judgement which convinced him of the validity of what he was doing. He must have been sensitive to things that other people were blind to and was able to tell from the logical properties of his equation and things relating to it that it had a much better chance of being true than the other possibilities he had thought of. Actually he put it more strongly than this, namely that he *knew* he was right just from the form of the equations in the context that he was working. I think this also points out the fact that aesthetic judgements vary so much from person to person. Different people do not have the same views as to which things are beautiful; but some people seem to have better judgements in a particular context. Here is a place where you can, in a sense, measure how good aesthetic judgements are. Normally these judgements seem so subjective and untestable. How often one hears critics discussing the arts: one person says something is wonderful and another says the same thing is terrible! Well now we have an area in which you can, in a way, test somebody's aesthetic judgement. Indeed it seems to me that this is, perhaps, an area for research in itself. That would be a very intriguing possibility. It would, of course, depend upon people like Dirac being able to repeat their successes. To make one brilliantly successful judgement like this but never to do it again would provoke the thought that the chance element might be strong. But Dirac did seem to have the habit of doing this sort of thing several times, and so there must be far more to it than just chance. I could go on to mention other things in theoretical physics which have a special beauty of their own. Maxwell's equations are a very good example—and Maxwell, also, was able to repeat many successes. Of course, quantum theory and general relativity are, each in their own way, things of great beauty. But quantum theory was the product of many minds and many differing aesthetic judgements must have gone into its discovery. General relativity was the product essentially of one mind and is correspondingly more of a unity.

Let me now turn to the question of what it is that constitutes elegance. It is naturally a difficult question, but I think Professor Atiyah made some throw-away remark about this in his talk, mentioning that elegance is more or less synonymous with simplicity. I think that there is considerable truth in this. Elegance and simplicity are certainly things that go very much together. But nevertheless it cannot be quite the whole story. I think perhaps one should say it has to do with *unexpected* simplicity, where one imagines that things are going to be complicated but suddenly they turn out to be much simpler than expected. It is not unnatural that this should be pleasing to the mind. One was expecting to have to do a lot of work—and suddenly there is much less work to do than had been anticipated. Quite reasonably this seems pleasurable. Perhaps this is over-simplified, but I think there may be something in it. Of course, simplicity by itself is not enough. Nobody would say that the equation $2 + 2 = 4$ is particularly beautiful. I cannot remember back far enough to the time when I first came across this particular theorem and I do not remember any stage in my life when I regarded $2 + 2 = 4$ as anything especially beautiful. On the other hand, the fact that $2 + 2 = 2 \times 2$ had, I think, a certain appeal to me at some time in the distant past. Of course, when one first comes across something it may have a certain beauty which later disappears when it becomes too familiar—indeed, it may become boring after a while. Now I can remember a time when, at quite an early age, I noticed that $3 + 1\frac{1}{2} = 3 \times 1\frac{1}{2}$, so I mentioned this fact to my brother. He was a little older than I was and considerably more able at mathematics, and he produced, after a little calculation, the fact that any two numbers whose reciprocals sum to unity have the property that their sum is equal to their product. He then showed me how to arrive at the result algebraically and I was fascinated by this. It was the first time I had ever seen how results could be simplified by means of algebra and it was extremely mysterious and beautiful to me. I suppose again it is one of those cases where the beauty arises because things turn out to be expressible in a much simpler way than had been anticipated. And at that age it had seemed a fantastic thing to be able to collect together all the pairs of numbers whose sum equals their product into a single formula which seemed quite different yet surprisingly simple. So I think that unexpected simplicity was quite a significant thing here also. And there is the element of connecting together different ideas that one had not expected to be related. This is something that frequently occurs in mathematics and it has a special beauty of its own. I think the principle of beauty in unexpected simplicity applies quite generally to such situations too. Two problems which seem unrelated can both involve one in a certain amount of work; but as soon as it is realised that the two problems are really basically the same problem, the amount of work may be halved—and the pleasure of unexpected simplicity is consequently felt!

It is, of course, possible to give many examples of unexpectedly simple proofs of theorems. There is no doubt that such proofs have an especial elegance of their own. One of my favourites is the proof of Pythagoras' theorem for which the diagram is given in Fig. 1. Since the three triangles in the diagram are similar, their areas are in proportion to the squares on their respective hypotenuses. The result follows. There are many other simple proofs of geometrical theorems. It so happened that in my own work I once needed to use Ptolemy's theorem which expresses the condition for four points to be concyclic in terms of the distances between them. After struggling with complicated arguments of my own devising I looked up Ptolemy's original proof which I had forgotten. I was amazed by the beauty and simplicity of the argument, which just involved drawing one unexpected construction line.

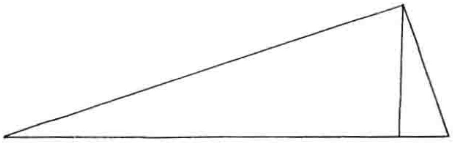


Fig. 1

I should like to stick with geometry for a while and attempt to explore this question of beauty in simplicity further. Consider the square lattice in Fig. 2, for example.

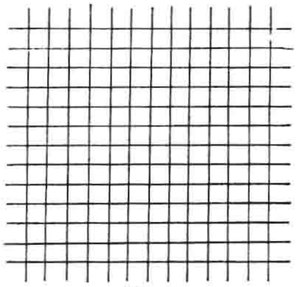


Fig. 2

There is no doubt that it is something simple. But I do not really feel that there is much beauty about it. As a pattern it is just boring. It may just be a question of familiarity, of course. We see square or rectangular lattices so much in our everyday lives that the element of unexpectedness is completely removed. But now consider Fig. 3. It is a photograph of a window in a 14th century mosque in Cairo. The illustration is taken from H. Weyl's "Symmetry." I think people will agree that

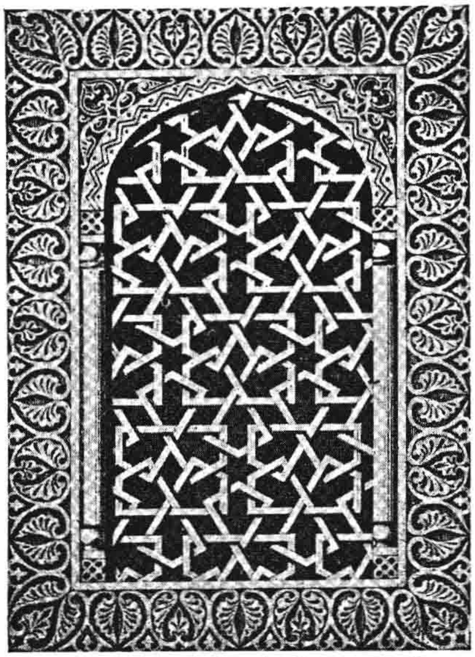


Fig. 3. Reproduced from Speiser, A., "*Theorie der Gruppen von endlicher Ordnung*," 3 Aufl., 1924, with permission of the publishers, Springer Verlag.

this has considerably more aesthetic appeal than the square lattice and yet it is much more complicated. The question is: why should it have more appeal? Now I am not sure whether the appeal can be regarded as something just mathematical, or whether there is an essentially different kind of aesthetic appreciation also involved. The design has some rather unexpected features. There are straight lines continuing across which make up a succession of interlacing trefoil knots. The unexpectedness of the pattern may well be something which adds to its appeal, I suspect. But I do not see that one could really call the pattern simple. It may be that *symmetry* itself could generally be put under the heading of unexpected simplicity on the other hand. It is only a tentative hypothesis, but the idea would be that with a symmetrical configuration one can appreciate the whole by examining only a portion of it. There is less work for the mind to do to comprehend the configuration than there would be without the symmetry.

Let me turn now to Fig. 4. Really the reason for my choice of topic for this whole lecture was to show this pattern! I had often doodled by fitting together limited configurations of pentagons and similar shapes but I had never found a good rule for continuing such patterns indefinitely. However, recently I wanted to design something interesting for someone who was in hospital to look at and I realised that there was a certain definite rule whereby one could continue such a pattern to arbitrary size. The pattern never repeats itself—in the sense that there is no period parallelogram. Nevertheless, the rule that forces this pattern is a purely local one. Notice that the entire configuration is made up of four different shapes: pentagon, five-pointed star, a kind of three-pointed half-star and a diamond. There are rules about which of these shapes are to fit next to which others. Perhaps the easiest way to state things is to modify the shapes slightly and make a kind of infinite jigsaw puzzle out of them. I have depicted in Fig. 5 a way that one might do this. There are three different pentagon pieces now, making six pieces in all. Of course the exact shape of the notches is not significant, but the arrangement of the notches is. The puzzle would be to cover the entire plane with these six types of piece, where there is to be an infinite supply of each piece. It turns out that in a certain "finite" sense the assembly of the pieces is unique—even although the pattern does not repeat! The entire completed pattern is not, strictly speaking, unique. But given any two complete patterns, any *finite* portion of either one is to be found somewhere in the other (infinitely many times over, in fact). I shall not go into the proof of this here, but roughly speaking the idea is to show that by fitting pieces together one is inevitably forced to build a larger copy of one of the original pieces. Proceeding by induction one shows that the entire

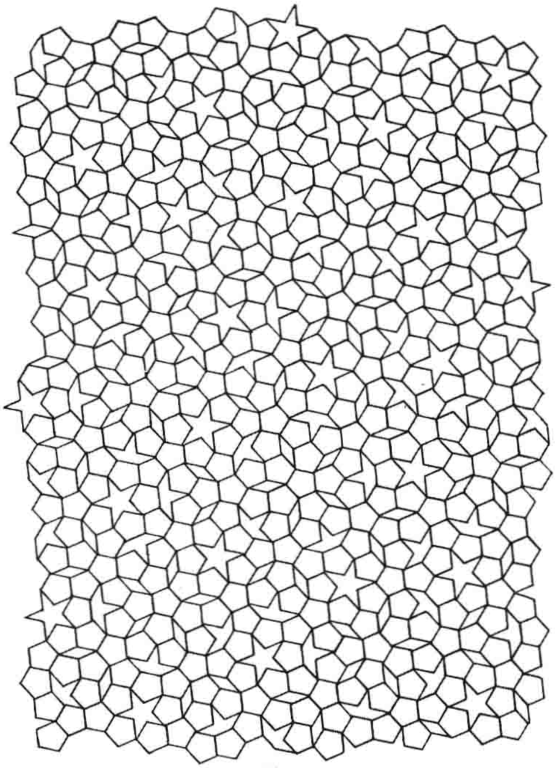


Fig. 4

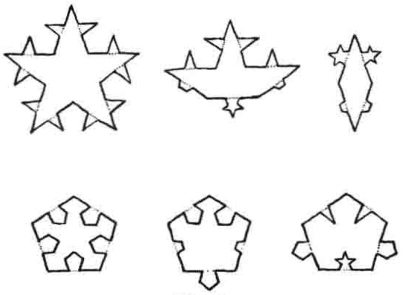


Fig. 5

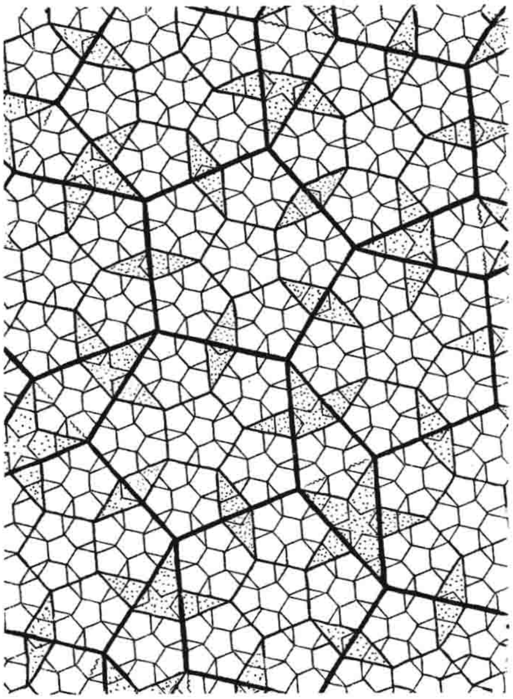


Fig. 6

pattern must be constructible by a process of subdivision of pieces into smaller pieces followed by an expansion, so that the subdivisions come back to the size of the original pieces. The way in which this process applies to the pattern of Fig. 4 is illustrated in Fig. 6.

I feel that the pattern of Fig. 4 has a certain aesthetic appeal which seems to be independent of the features I mentioned above. As one simply stares at the pattern certain regularities seem to jump out. There are a great many regular decagons which tend to overlap at various places. Each decagon is surrounded by a ring of pentagons. There are larger rings of various kinds too, such as some consisting of ten diamonds. Also there is a certain duality possessed by the pattern, although this is not obvious. And there is a simple feature that is rather surprising which I had not realised the pattern possessed before drawing it out. Each line segment in the picture continues to a straight line which contains infinitely many other line segments of the pattern. Things line up in a surprising way. The appeal of this pattern would seem to have something in common with the appeal of the mosque window depicted in Fig. 3. But, in addition, there seems to be something of aesthetic value in the fact that the pattern almost repeats and then does not quite. One does not have to know the mode of construction of the pattern in order for it to have an aesthetic appeal. I think that the suggestion I made about "unexpected simplicity" lying at the root of aesthetic judgements can at best be an over-simplification. It is difficult to see how such ideas will apply here.

Let me attempt to draw a certain moral. Perhaps it is not a moral exactly, but it seems that there is something of an example of the way in which research often proceeds in all this. One often becomes fascinated by a problem even although it has no particular relevance to other things one may be doing. And when one is fascinated, the internal aesthetics of the thing will drive it along. Sometimes it turns out that what one has been doing can be used for something else, but sometimes not at all. That is one of the nicest things about mathematics. When you have something pretty, then this, in itself, can be regarded as its justification. You can publish it if you wish, without caring whether anybody else can find a use for it. So often it must be the case that people just do things for their own sake, merely for the enjoyment of doing them, and this enables progress to be made that could not be made otherwise. Sometimes it is not exactly for enjoyment. It may be more a kind of obsession that takes over. Let me mention something of this kind which happened to me once. I was writing a paper on something in general relativity, where I had given a number of conditions for asymptotic flatness for a space-time. Then there was an extra assumption, which was topological in nature. Now this extra assumption could easily have been included as another condition. For the kind of physical situation I was concerned with this extra assumption was obviously no physical restriction and no-one would object to having it included with the conditions for asymptotic flatness. It was not really something I needed to worry about at all. Nevertheless it seemed to me that it was probably logically dependent on the other conditions and one could get rid of it. The thing kept nagging at me and it held up the paper for quite a while. I kept thinking I must be able to prove this. Then I thought I had found a proof and I started to put a sketch of it in an appendix. But presently I discovered there was a mistake in the argument, so I had to stop the writing again. I realised there were some subtleties I had not thought of before and I began thinking about new types of argument to deal with them, all the time thinking I was wasting my time completely. Finally a rather complicated and untidy argument emerged which I outlined in the appendix to my paper. The whole thing was insignificant from the point of view of what I was trying to do in the paper, but it would never let me go; I could never stop thinking about it until I felt I had sorted the matter out. And then many months later, the same kinds of argument I had been using in this particular result turned out to be very useful in a different context which I had not been thinking about at all. Suddenly they could be used in a way which was much more significant and important within relativity theory. Now this is an example of a case where one never really knows when one is wasting one's time. I have often noticed this phenomenon. In this particular case I certainly thought I was wasting my time all the time I was working on the problem I mentioned. But whether one can call this desire to sort out that problem an aesthetic matter, or just an obsession, I am not sure!

Finally I would like to say something about symmetry in a set of equations, in this case Maxwell's equations. Writing these out in full and taking the case of zero sources we have

$$\frac{\partial G}{\partial y} - \frac{\partial F}{\partial z} + \frac{1}{c}\frac{\partial H}{\partial t} = 0 \qquad \frac{\partial L}{\partial y} - \frac{\partial K}{\partial z} - \frac{1}{c}\frac{\partial E}{\partial t} = 0$$

$$\frac{\partial E}{\partial z} - \frac{\partial G}{\partial x} + \frac{1}{c}\frac{\partial K}{\partial t} = 0 \qquad \frac{\partial H}{\partial z} - \frac{\partial L}{\partial x} - \frac{1}{c}\frac{\partial F}{\partial t} = 0$$

$$\frac{\partial F}{\partial x} - \frac{\partial E}{\partial y} + \frac{1}{c}\frac{\partial L}{\partial t} = 0 \qquad \frac{\partial K}{\partial x} - \frac{\partial H}{\partial y} - \frac{1}{c}\frac{\partial G}{\partial t} = 0$$

$$\frac{\partial H}{\partial x} + \frac{\partial K}{\partial y} + \frac{\partial L}{\partial z} = 0 \qquad \frac{\partial E}{\partial x} + \frac{\partial F}{\partial y} + \frac{\partial G}{\partial z} = 0$$

$\vec{E} = (E, F, G)$,	$\vec{H} = (H, K, L)$,	$\vec{h} = (x, y, z)$
electric field	magnetic field	position vector

It would seem that the symmetry of these equations and the aesthetic appeal that this symmetry generated must have played an important rôle for Maxwell in his completion of the equations. Now these equations may be re-expressed in the familiar form

$$\text{curl}\,\vec{E} + 1/c\,\vec{H} = 0 \qquad \text{curl}\,\vec{H} - 1/c\,\vec{E} = 0$$

$$\text{div}\,\vec{H} = 0 \qquad \text{div}\,\vec{E} = 0$$

or, equivalently, in the notation of world-tensors or of differential forms:

$$\nabla_a F_{bc} + \nabla_b F_{ca} + \nabla_c F_{ab} = 0, \quad \nabla^a F_{ab} = 0$$

i.e., $$d\phi = 0, \quad \delta\phi = 0.$$

Up to now we have equations relating six real functions of four real variables. But the equations can be re-expressed in spinor terms in the single formula

$$\nabla^{AC'}\phi_{AB} = 0,$$

where ϕ_{00}, $\phi_{01} = \phi_{10}$, ϕ_{11} are three complex functions of the four real space-time variables. The number of equations is reduced from eight to four. I shall not attempt to explain this formula, but merely state that by writing Maxwell's equations in this way one can see that they form part of a pattern which includes other physical fields as well. Then by re-expressing the space-time coordinates in a completely different form, the Maxwell field can be described by a single complex (analytic) function of four complex variables $f(W_\alpha)$. The field equations have now completely disappeared and are replaced by the single homogeneity relation

$$f(\lambda W_\alpha) = \lambda^{-4} f(W_\alpha).$$

If it were not for the initial symmetry (and for the appeal that the equations must have had because of the symmetry) this kind of formulation would never have been possible. I am sure that it could not have been evident to Maxwell that one could eventually have had this type of formulation. Nevertheless I feel that there is some considerable significance in the fact that it exists because, for example, the complex numbers which arise fit in with quantum theory in interesting ways. I do not want to go into it here. The point is that one often does not know *why* symmetry has an appeal, but nevertheless the fact that it is there enables something deeper to emerge eventually, but not perhaps for many, many years, possibly centuries, later.

Bibliography

Penrose, R., *Proc. Roy. Soc. (Lond.)*, 1965, **A284**, 159.
Penrose, R., and MacCallum, M. A. H., *Physics Reports*, 1973, **6C**, 242.
Weyl, H., "Symmetry," Princeton University Press, 1952.

65

SINGULARITIES IN COSMOLOGY

Confrontation of cosmological theories with observational data.
Ed. M. S. Longair, pp. 263–72. Springer, 1974

Here, using the TIP and TIF procedure of Chapter 50, I was groping for a good mathematical characterization to distinguish the singularity at the big bang from (the time-reverse of) those in black holes, realizing that such a distinction was needed for the second law of thermodynamics, and that this could be clear cut if the strong form of cosmic censorship referred to here holds true. I had not yet thought of using the Weyl tensor to provide the needed distinction, however. See also Chapters 91, 96, 104, and 270.

750 Singularities in cosmology. *Confrontation of cosmological theories with observational data.* Ed. M. S. Longair, pp. 263–72

SINGULARITIES IN COSMOLOGY

ROGER PENROSE
Mathematical Institute, Oxford, U.K.

Abstract. Singularities in space-time can be broadly divided into three classes: past-spacelike (in white holes or the big bang), timelike (naked singularities) and future-spacelike (in black holes or the final recollapse). In a closed Universe, if a simple restriction is made to eliminate timelike singularities, the inference may be drawn that the topology of the Universe is unchanging with time. Thermodynamical considerations lead one to infer that the final singularity of recollapse must differ markedly in structure from the initial big bang. This may plausibly be related to the existence of black holes and the presumed non-existence of white holes.

In this lecture I shall confine myself to making general qualitative remarks about singularities. The previous speaker, Prof. Lifshitz has given an excellent account Belinskii *et al.*, 1974) of the very fine and detailed work (Khalatnikov and Lifshitz, 1963; Belinskii *et al.*, 1970, 1972) carried out over a number of years by the Soviet school. I do not feel able to add to that here.

For the sake of simplicity I shall also confine my remarks to the case of closed universe models only. This is not intended to reflect any bias on my part as to whether I believe the Universe to be open or closed in fact. I have no strong feelings on the matter. It is merely that the statements of the results that I shall be concerned with are much more clear-cut in the case of closed universe models than open ones. Generalizations to the case of open universes are certainly possible and are treated to some extent in my other lecture (Penrose, 1974).

Let me begin by reviewing the main singularity theorem from which we can infer the existence of some form of singularity in a general closed universe model:

THEOREM (Hawking and Penrose, 1970): *A space-time which*

(i) *contains no closed timelike curves.*

(ii) *satisfies Einstein's equations (without cosmological term) and the energy condition* ($\varrho + p_i \geqslant 0$, $\varrho + \sum p_i \geqslant 0$),

(iii) *is sufficiently general (i.e.* $t_{[a}R_{b]cd[e}t_{f]}t^c t^d \neq 0$ *somewhere along each timelike or null geodesic,* t^a *being the tangent vector), and*

(iv) *contains a closed spacelike hypersurface,*

cannot be geodesically complete in all timelike and null directions.

The concept of geodesic incompleteness is not quite the same as that of a physical singularity. One would expect that in the actual universe, the 'reason' for the geodesic incompleteness would be that space-time curvatures become so large that the local physics becomes drastically affected – to the extent that the normal ideas of space and time might break down. But the theorem says nothing about curvatures becoming large. That is its main weakness. Nevertheless, if one applies the theorem to a universe model which is maximally extended, then the association of geodesic incompleteness with *some* form of physical singularity seems reasonable – even if in particular cases, the singularity might be, say, of the 'conical' type with curvatures remaining finite right up to the singularity.

The theorem does not enable us to locate the singular points – it is very negative in this respect. Nor does it supply a definition of something that we could actually call a singular point. In fact various alternative definitions of singular points have been given, each with its own particular merits (cf. Hawking and Ellis, 1973). The definition I shall use here is basically that given by Geroch *et al.* (1972). I shall apply the definition to universe models which are closed in the sense of having no points at infinity. Thus I assume, for simplicity, that no timelike curves exist with infinite proper length.

Let me first indicate in very crude and oversimplified terms how we might expect to classify singularities broadly into three groups:

(i) past-spacelike (big bang, white hole)
(ii) timelike (naked singularity)
(iii) future spacelike (black hole, final singularity).

I am including 'null' under the heading of 'spacelike' here; it is not hard to make a precise distinction but I shall not bother with it. The essential feature of a past spacelike singularity is that it supplies a past singular end-point to an otherwise past-endless timelike curve. A future spacelike singularity supplies a future singular end-point to an otherwise future-endless timelike curve. A timelike singular point may be thought of as supplying *both* a past end-point to a past-endless timelike curve and a future end-point to a future-endless timelike curve. (However, I shall not quite define things this way.) Two past-endless timelike curves γ, η are deemed to have the same (singular) past end-point if and only if they have the same futures (Figure 1). Likewise, two future-endless timelike curves γ', η' are deemed to have the same (singular) future end-point if and only if they have the same pasts.

I should like to indicate why I feel that this method of identifying singular points is rather natural from the physical point of view. Let us consider the big bang singularity of the normal cosmological models. If we think of the initial singularity as a *single* point then we have the unnatural situation that because of the existence of particle horizons, this point gives rise to an infinity of causally disconnected regions at the next instant. It seems more natural to think of the singularity as a three-dimensional spacelike surface. (This picture can be obtained by a conformal rescaling of the other one (cf. Penrose, 1968).) Each point of this surface has a distinct domain of influence, the 'future' of that point (Figure 2). The above definition of singular points leads naturally to such a three-dimensional description of the big bang singularity. Each point of the singular set is directly associated with the region of space-time that it can influence.

Let us see why, from this point of view, the normal black hole ($r = 0$) singularity of the Schwarzschild solution must be regarded as future-spacelike. The situation is depicted in Figure 3. All future-endless curves inside the event horizon must hit the singularity and no timelike curve can leave it. This singularity turns out also to be three-dimensional like that of the big bang. Timelike curves entering the singularity

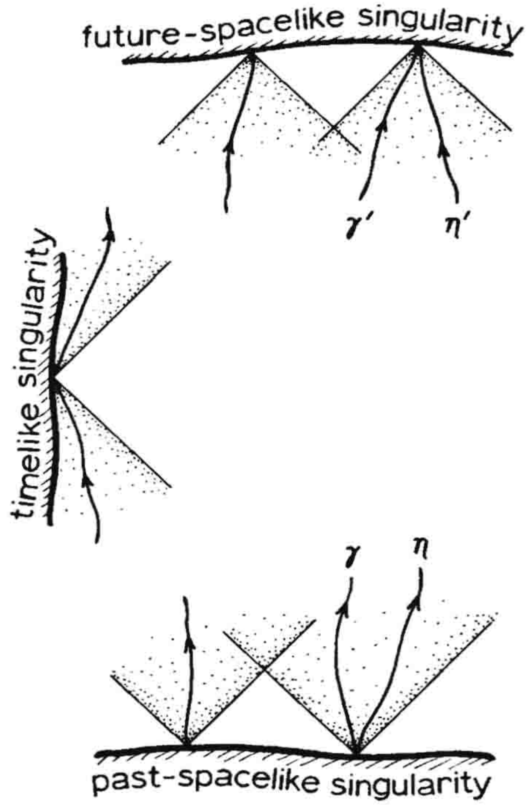

future-spacelike singularity
γ'
η'
timelike singularity
γ
η
past-spacelike singularity

Fig. 1.

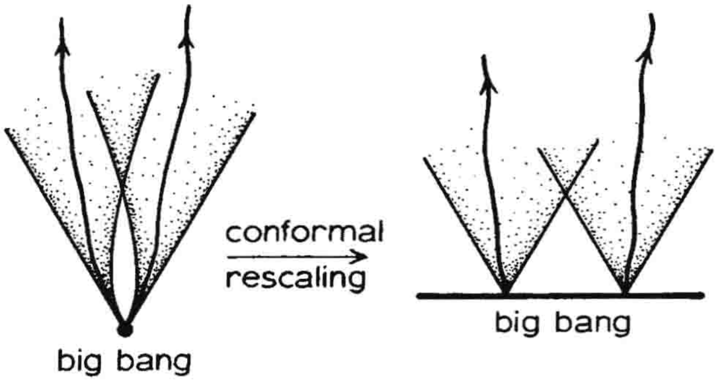
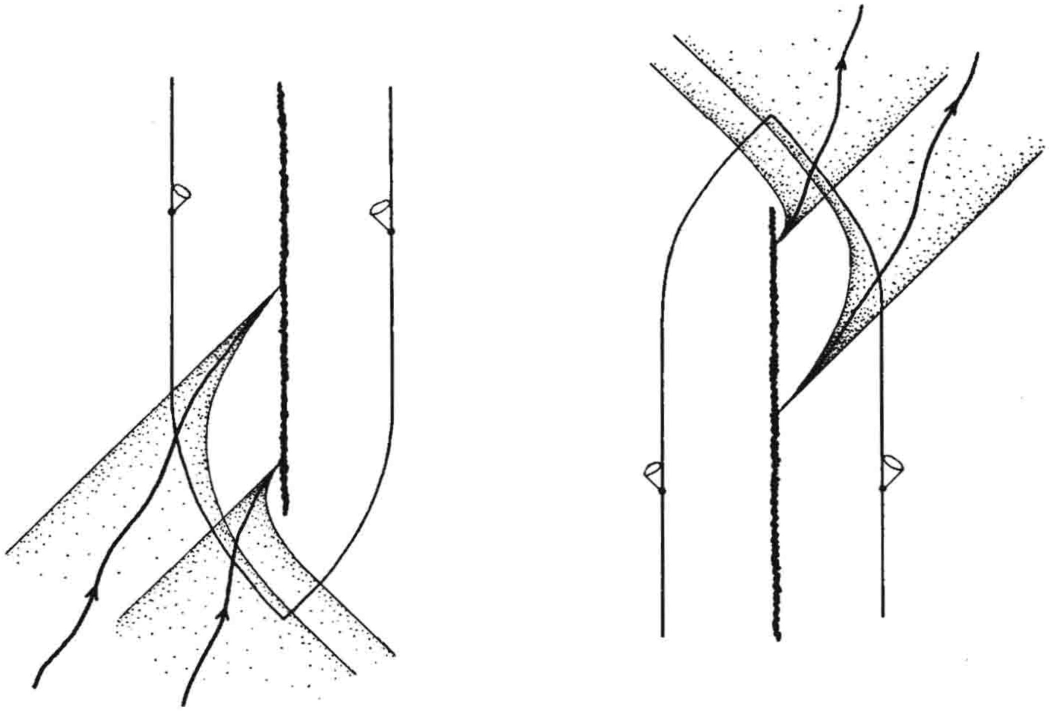

big bang
conformal
rescaling
big bang

Fig. 2.

from different directions or at different 'times' will have distinct pasts. The space-like character of this singularity is perhaps at variance with ones initial intuition since there is clearly a sense in which it persists with time. Nevertheless its local spacelike nature seems inescapable. I shall indicate later a possible criterion for distinguishing the black hole type of singularity from the more cosmological sort. The situation for



black hole

white hole

Fig. 3.

white holes is just the time-reverse of that for black holes. Again the singularity is three-dimensional and spacelike – at least, in the spherically symmetrical case.

The situation arising with a naked singularity can be formulated as follows. Suppose there is a point p in the space-time and a past-endless timelike curve γ whose future is contained in the future of p (Figure 4). Since γ is past-endless and since points at infinity have been excluded, we must assign a singular past end-point q to γ. But the future of q (being the future of γ) lies in the future of p. Thus the singular point q itself may be reasonably thought of as lying to the (causal) future of p. Thus q is both in the future of some observer (say p) and in the past of another (the points of γ). So we may regard q as a *timelike* singular point. Indeed, since γ lies to the future of p, we may choose some observer (timelike curve) who starts at p and travels to some point r on γ. Then the singular point q lies to his future when he is at p and to his past when he is at r. This is the situation of a *naked singularity*. We may, for example, envisage a gravitational collapse taking place, where the state is initially non-singular but where the system evolves into a singular configuration, the resulting singularity being, unlike that in a black hole, visible to observers at large distances. Such singularities are called naked, and it is customary to rule out their existence by a hypothesis: the hypothesis of *cosmic censorship* (cf. Penrose, 1974). In fact, it turns out that if we adopt the form of this hypothesis that excludes singular points q of just the type considered above, then the space-time must be *globally hyperbolic*. This means (cf. Geroch, 1970) that the entire space-time can be evolved from a single Cauchy hypersurface – indeed that the topology of the Universe must be unchanging for all time. For a closed universe (no points at infinity) we may regard the condition of global hyperbolicity as equivalent to cosmic censorship. The condition is actually time-symmetrical: exactly the same condition arises if, in the above, q had been taken as the singular future end-point of a future-endless timelike curve γ' and r as a point whose past contains the past of γ'.

It is perhaps hard to visualize how the topology of the Universe can be unchanging in a situation where there are black holes, and possibly also white holes, where new black holes may be forming, perhaps old white holes disappearing, and where distinct black holes may be congealing into one. In fact, the Cauchy hypersurface must have the property that it passes 'underneath' all the black hole singularities and 'above' all the white hole singularities. The situation is indicated in Figure 5. The Cauchy hypersurface remains spacelike inside the black holes' horizons owing to the 'tipping' effect on the light cones.

It should be observed that our hypotheses have not ruled out the possibility of white holes. Indeed, it is difficult to distinguish, on purely qualitative grounds, between a white hole singularity and the big bang itself. An argument has been given by Zel'dovich (1974) to the effect that white holes ought spontaneously to decay in a very short period of time owing to particle creation effects at or near the singularity. It should be emphasized that the distinction between black and white holes which seems to occur here is closely bound up with the statistical nature of time-directivity and to the asymmetric behaviour of radiation. We recall that radioactive nuclei are observed to *decay*,

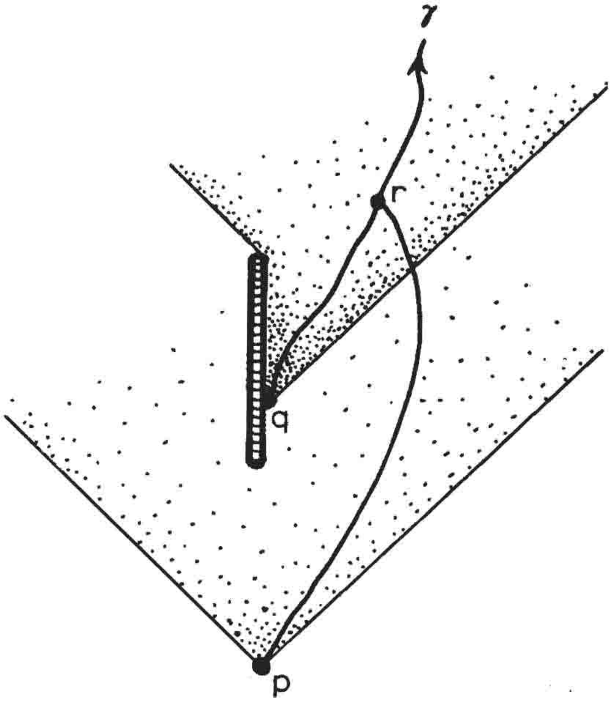


Fig. 4.

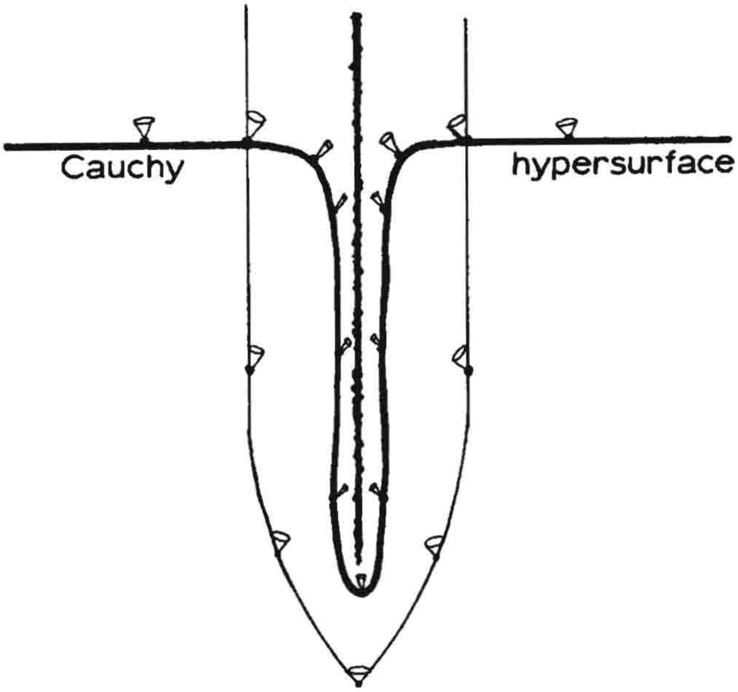


Fig. 5.

accompanied by the emission of radiation to infinity. But we do not expect such nuclei spontaneously to assemble themselves while absorbing radiation sent in from infinity. The local physical laws are the same in each case, but the type of boundary conditions that one allows are quite different for the two cases.

When the implications of this distinction between past and future are examined in the setting of a closed universe, one is inevitably confronted with the question as to whether there is an essential difference between the past and future space-time singularities. The 'blame' for any statistical difference in the boundary conditions in the past and future gets pushed back into the singularities themselves. One possibility for the distinction between past and future is that the very slight violations of time-reversal invariance in local physics which *are* observed (Casella, 1968) might play some crucial and greatly magnified role in the high curvature regimes which neighbour the singularities themselves. Another possibility that is sometimes considered (although I personally find it hard to take too seriously – particularly when examined in relation to black holes) is that the statistical time's arrow will somehow reverse itself when the Universe – assumed closed – reaches maximum expansion. Setting aside such possibilities, and still retaining, for simplicity, the picture of a closed universe, we are driven to ask how a possible difference in singularity structure between the big bang and the final recollapse might in some way be related to the statistical time's arrow.

We can look at this question in a somewhat different way. Let us imagine that some appropriate definition of entropy can be found which refers to the whole Universe and includes gravitational effects (i.e. it does not refer just to systems of particles on a given curved background, (cf. Tolman, 1934), but the curvature itself must contribute appropriately to the entropy). In the early stages of the big bang one has a picture of something resembling thermal equilibrium – at least that is how calculations are done. So we envisage that the entropy (ignoring possible gravitational curvature contributions) is somehow almost at its maximum for the given 'size' that the Universe has at that time. As the Universe expands, the entropy increases, but lags behind the possible theoretical maximum that might be allowed for a universe of that size (Figure 6). Then the Universe recontracts and the theoretical maximum entropy is reduced with it. But since the entropy of the Universe has been increasing all the time, the Universe cannot reach a state resembling the compact system in thermal equilibrium in which it appeared to start out. Somehow the gravitational irregularities – apparently contributing positively to the entropy – must have grown. The model is now presumably riddled with black holes, or something similar which can be defined for a closed universe. One seems driven to some sort of concept of black hole entropy – or, at least, some sort of positive entropy residing in curvature irregularities.

In this connection one is reminded of the concept, due to Beckenstein (1973) (and the related ideas due to Bardeen *et al.*, 1974) of a *black hole entropy* S_{bh} defined by the formula

$$S_{bh}=\frac{\eta k c^3}{\hbar G}A,$$

where η is some numerical constant of order unity, k is Boltzman's constant, and

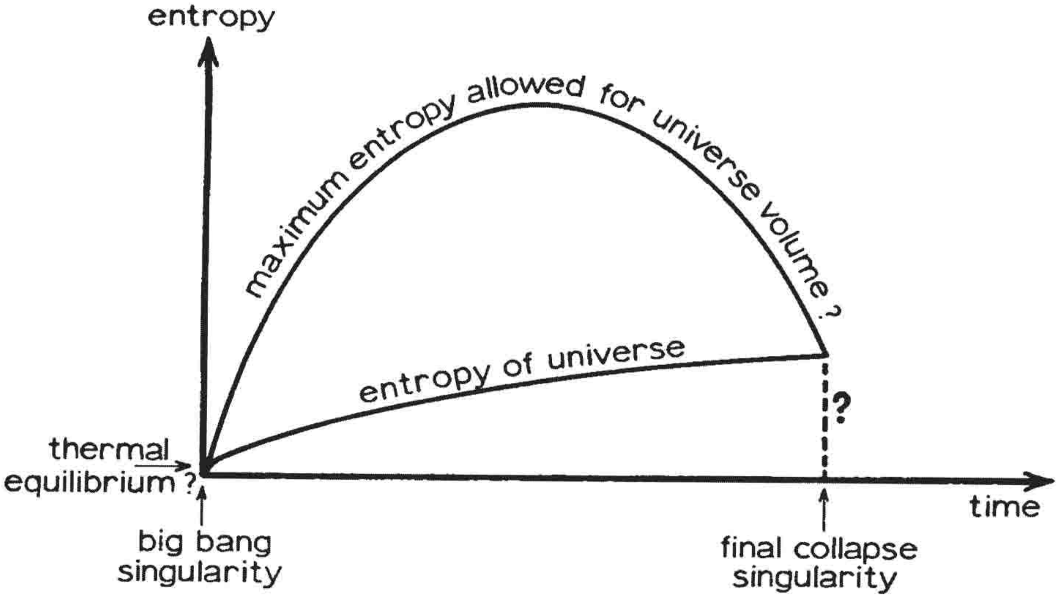


Fig. 6.

where A is the surface area of the black hole. This definition is intended for use only in asymptotically flat space-times and the formula has been the subject of some controversy. I do not propose to take sides on the issue here. I only mention it as one possible line of approach which has some bearing on the difficult and somewhat nebulous questions that I am attempting to raise.

In Figure 7 I have tried to depict how an irregular closed universe full of black holes might look from the space-time point of view, where I am assuming that the cosmic censorship principle holds and that the model is free of white holes. But from the point of view of the topological or causal structure, all the 'stalactites' which represent black hole singularities could be straightened out and the final singularity might appear indistinguishable from the initial singularity. How, in fact, are we to make precise, in mathematical terms, the very real distinction between the initial and final singularities? I shall just give one possible criterion, for what it is worth, and then leave the matter there.

One virtue of global hyperbolicity is that any two points with a timelike separation have a well-defined maximum time interval between them, this being the length of a maximal timelike geodesic connecting them (cf. Penrose, 1972). This time interval is actually a continuous function of the pair of points. We can extend this concept to apply also to the singular points. We can then ask, for each *given* past singular point, what is the maximum time-interval to future singular points. We find that this maximum varies little as the past singular point is varied. In a corresponding way, we can ask for the maximum interval from past singular points to some *given* future

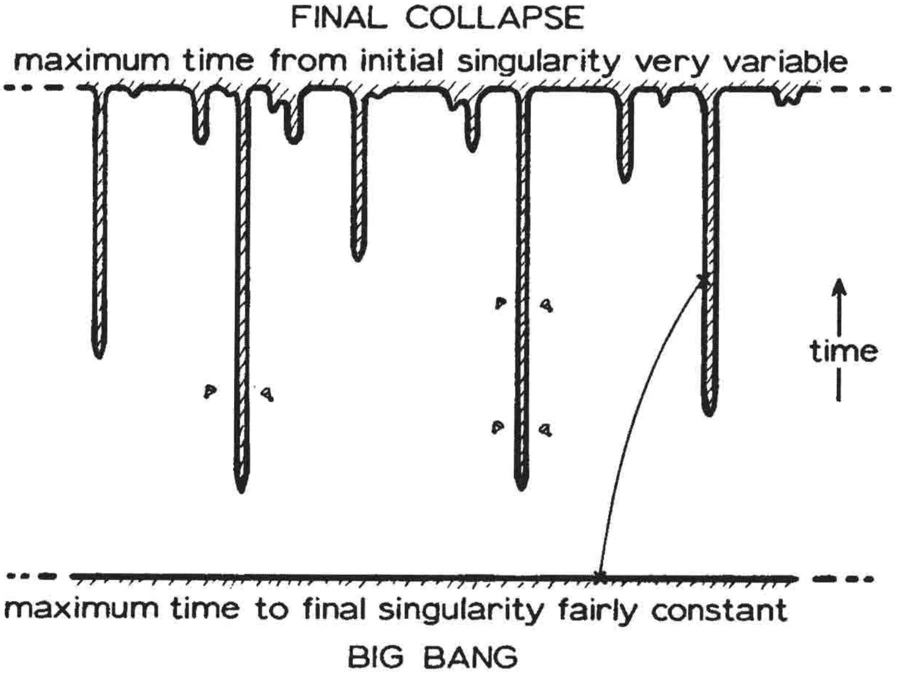

FINAL COLLAPSE
maximum time from initial singularity very variable
p
q
p
q
p
q
time
maximum time to final singularity fairly constant
BIG BANG

Fig. 7.

singular point. We now find that this maximum varies very greatly as the future singular point is varied. The interval will be small for a singular point on a black hole singularity which occurs early in the Universe's history but it will be large for a singular point which only results from the final recollapse of the entire universe. How one could precisely relate such a time-asymmetry in the singularity structure to the statistical asymmetry of the details of the physical universe is a very open question, however.

References

Bardeen, J. M., Carter, B., and Hawking, S. W.: 1974, to appear.
Beckenstein, J. D.: 1973, *Phys. Rev.* **D7**, 2333.
Belinskii, V. A., Khalatnikov, I. M., and Lifshitz, E. M.: 1970, *Adv. Phys.* **19**, 523.
Belinskii, V. A., Khalatnikov, I. M., and Lifshitz, E. M.: 1972, *Soviet Phys. JETP* **62**, 1606.
Belinskii, V. A., Khalatnikov, I. M., and Lifshitz, E. M.: 1974, this volume p. 261.
Casella, R. S.: 1968, *Phys. Rev. Letters* **21**, 1128.
Geroch, R.: 1970, *J. Math. Phys.* **11**, 437.
Geroch, R., Kronheimer, E. H., and Penrose, R.: 1972, *Proc. Roy. Soc. London* **A327**, 545.
Hawking, S. W. and Ellis, G. F. R.: 1973, *The Large Scale Structure of Space-Time*, Cambridge Univ. Press.
Hawking, S. W. and Penrose, R.: 1970, *Proc. Roy. Soc. London* **A314**, 529.
Khalatnikov, I. M. and Lifshitz, E. M.: 1963, *Adv. Phys.* **12**, 185.
Penrose, R.: 1968, in C. M. DeWitt-Morette and J. A. Wheeler (eds.), *Battelle Rencontres*, Benjamin, New York.
Penrose, R.: 1972, *Techniques of Differential Topology in Relativity*, S.I.A.M., Philadelphia.
Penrose, R.: 1974, in C. M. DeWitt-Morette (ed.), 'Gravitational Radiation and Gravitational Collapse', *IAU Symp.* **64**, 82.
Tolman, R. C.: 1934, *Relativity, Thermodynamics and Cosmology*, Clarendon Press, Oxford.
Zel'dovich, Ya.: 1974, *Proc. IAU Symp. 64 on Gravitational Radiation and Gravitational Collapse*, Warsaw, 5–8 September 1973.

DISCUSSION

Zel'dovich: What changes occur in the case when $\Omega < 1$ and the Universe is open? In this case black holes also form but most of the geodesics do not end in these black holes but go to the future infinity.

Penrose: I discussed the question of open models in my contribution to Symposium No. 64 in Warsaw. The corresponding considerations for open models are similar to those for the closed models provided, in the former case, we consider points at infinity to be in the same category as singular points.

Misner: Suppose white holes are admitted. Consider the following possibility. The initial singularity contains a small white hole which will not explode out into the Universe until a late time. Before that time it has fallen within the horizon of a large black hole. Where now can the Cauchy hypersurface be fitted in?

Penrose: There is no difficulty about the Cauchy hypersurface entering the event horizon of the black hole and then leaving it again, since this hypersurface is spacelike. It can approach the singularity as closely as you like and then slip underneath it.

Misner: You referred to Zel'dovich's arguments in Symposium No. 64 to support the hypothesis that white holes may be avoided. The theoretical necessity for pair creation seems inevitable, but I do not see that this controls the epoch at which the white hole ejects matter into the external universe. Since time translations $t \to t - t_0$ in the external Schwarzschild field correspond to space translations along the past spacelike singularity in the white hole (Schwarzschild $r \to 0$), it would seem that some inhomogeneity feature of the pair creation process fixes the white hole explosion epoch, and Zel'dovich did not estimate such an inhomogeneity parameter.

Novikov: The Schwarzschild singularity is spacelike. Pair creation does not change this property. But pair creation near the spacelike singularity changes the initial conditions which determine the evolution of the space-time from the singularity to the future. It must change the sewing together of the singularity region and the external space. (The external space is in the absolute future from the singularity.) It can control the epoch at which the white hole ejects matter into the external space. My calculations are not finished yet, however.

Zel'dovich: I don't agree with Prof. Penrose when he said that the entropy of the Universe might decrease when a black or white hole forms. The entropy of the Universe increases with time and this law does not contradict the fact that we cannot measure the entropy of matter which has already fallen into balck holes. To put it in other words, we cannot consider a man to be destroyed when he has turned round the corner of a house and we don't see him.

66

TWISTORS AND PARTICLES: AN OUTLINE

Quantum theory and the structures of time and space. Eds L. Castell, M. Drieschner, and C. F. von Weizacker, pp. 129–45. Carl Hanser Verlag, 1975

This article provides a brief outline of *twistor particle theory* whereby the momentum and angular momentum of a *massive* particle is described by adding together the corresponding quantities for *n* massless particles, each of which is expressed in twistor terms. The freedom in doing this gives the *n*-twistor internal symmetry group, the case $n = 2$ being treated explicitly here. Brief earlier indications of this kind of idea are on p. [308] of Chapter 52.

762 Twistors and particles: An outline. *Quantum theory and the structures of time and space.* Ed. L. Castell, M. Drieschner, and C. F. von Weizacker, pp. 129–45

Twistors and Particles: an Outline

by Roger Penrose
Mathematical Institute,
Oxford.

1. Introduction

The object of twistor theory is to provide a new framework[1,2,3] for the mathematical description of basic physics. The conventional picture of a background space-time composed of points is, accordingly, to be replaced by a different one in which new fundamental entities, the twistors, take over the primary role. At first, twistor theory provides merely a reformulation: the normal space-time description is just translated into an alternative, but equivalent, mathematical framework. But the new framework suggests different directions in which to proceed, from the ones which might seem natural in a more conventional theory. In particular, twistors lead to a novel viewpoint as to the nature of a quantized space-time.[4,5] Also, a possible new approach to the classification of fundamental particles and their interactions is provided, a comprehensive scheme for the calculation of scattering amplitudes (hopefully free of the usual divergences) being part of the general programme.

A significant aspect of twistor theory is that the spaces encountered are invariably complex; and the appropriate functions on these spaces are holomorphic. These complex twistor spaces take over the role that the real spaces of position or momentum enjoy in the conventional formalism. Indeed, according to the twistor picture of things, the field of complex numbers lies nearer to the heart of physical laws than does the field of real numbers, not only at the level of quantum phenomena, but also in determining the very structure of space-time. It should be further remarked that the four-dimensionality of space-time is crucial in the twistor formalism; so one may perhaps regard the existence of a twistor description of Nature as providing some form of "explanation" for this four-dimensionality.

One may question, however, why the replacement of the real field by the complex field is to be regarded as a step in the right direction. Should we not, instead, be replacing both these continua by some sort of discrete structure so as to arrive, hopefully, at a form of combinatorial physics? In fact, such a long-term view of the ultimate nature of physical laws is to some extent a part of the twistor philosophy. There is a "rigidity" about holomorphic functions which is not to be found among smooth functions of real variables. In this sense, there is a much greater kinship between the complex and the combinatorial than between the real and the combinatorial.

As twistor theory stands at present, however, it cannot be said that this combinatorial aspect is very strongly brought out. Nevertheless, the general twistor programme receives some considerable motivation from the theory of "spin-networks", which is a model for the combinatorial description of a quantum geometry[6] that uses (discrete) quantum mechanical angular momentum as the primary physical concept from which the spatial relations are to be derived. Spin-network theory is based on the local isomorphism[7]

$$SU(2) \rightarrow O(3)$$

which is fundamental to the description of non-relativistic spin. In a corresponding way, twistor theory is based, in the first instance, on the local isomorphisms[7,8]

$$SU(2,2) \rightarrow O(2,4) \rightarrow C(1,3)$$

(C(1,3) being the 15-parameter conformal group of compactified Minkowski space-time).

The primary significance of C(1,3) lies in the fact that it is the symmetry group of free massless particles and fields[9] (of any helicity). With the introduction of rest-mass, C(1,3) symmetry is broken, the group having to be replaced by a subgroup of C(1,3) , namely the Poincaré group, describing the metric-preserving symmetries of Minkowski space-time. The twistors, which may be regarded as belonging to the representation spaces of SU(2,2) , are no less significant objects in the massive case than in the massless case. It merely becomes necessary that more than one basic twistor be introduced, at the outset, for the description of a (free) massive particle, where one twistor would have sufficed for a massless particle. In each case, the twistors describe the energy-momentum-angular-momentum structure of the particles and an "internal" symmetry group also arises, which represents the freedom of transforming the twistors among themselves while preserving this structure. For a single twistor description this "internal" group is U(1) , while for two twistors it is O(3) × E(2) (where E(2) is the group of orientation-preserving motions of the Euclidean plane and «~» means 2 fold covering). For three twistors[10], the group involves SU(3). The infinitesimal generators of these groups may be used to define twistor operators, the eigenvalues of which provide quantum num-bers (two of which are spin and mass) for the classification of particles.

With the introduction of gravity (and to a lesser extent electromagnetism as well) the twistor spaces become, in a sense, curved. The SU(2,2) and Poincaré symmetry groups no longer seem of direct relevance. Nevertheless twistor ideas still apply. The combination of curved space with the twistor quantization rules (that is, $Z^{\alpha}\bar{Z}_{\beta} - \bar{Z}_{\beta}Z^{\alpha} = \delta^{\alpha}_{\beta}$, where $\bar{Z}_{\alpha}$ is the complex conjugate of the twistor Z^{α}) seems to lead to a picture in which space-time points are "smeared out". The twistors themselves may be thought of as describing, on a classical level, the histories of spinning massless "particles", travelling with the speed of light (although individual twistors are not to be identified physically with massless particles). Space-time points may be defined as the "collision" points (intersections of the world-lines) of these "particles". On the quantum level, an uncertainty is introduced as to whether such a "collision" has taken place. In this way, the very concept of a space-time point becomes subject to quantum uncertainties.

In the conventional view of space-time, it is not normally regarded that our present modes of description of geometry should become inadequate until dimension of the order of 10^{-33} cm. are envisaged, this being the size at which the quantum fluctuations of the metric tensor would be large enough to produce alterations of the topology of space-time[11]. However, it seems likely that if twistors do turn out to provide a better formalism for microphysics than does the conventional space-time approach, then we shall begin to see the effects of this on the much larger scale of elementary particles (say 10^{-13} cm.). It has been argued that the agreement between quantum electrodynamics and experiment[12,13] shows that the normal space-time descriptions of nature must hold true down below 10^{-15} cm, but this should not be regarded as in any way contradicting the above statement. Newtonian theory does, after all, provide a very accurate description of planetary motion while still using (from the more satisfactory standpoint of genera relativity) the "incorrect" concepts of flat space-time geometry and gravitational force. Analogously, from the standpoint of twistor theory, the description of space-time in terms of points should be, likewise, "incorrect". Nevertheless, such conventional descriptions need not be seen to be in conflict with present observation.

My own attitude, that there is an absurdity involved in the normal space-time descriptions with respect to elementary particles, may be gleaned from the following picture. We wish to describe, say, a proton at a scale of, say, 10^{-23} cm. In the normal view, the continuous space-time picture is still eminently adequate, since the quantum fluctuations in the metric lie still ten orders of magnitude smaller than the scale considered. The points of space-time seem still as pointlike and indivisible as ever they were. But the proton itself appears as a barely discernable cloud[11] of enormous dimension, no less than ten orders of magnitude larger even than the scale considered. It seems quite incredible, in terms of such a picture, that a proton could behave as a single unit, with a very precise mass, charge and spin; and more incredible still that two such proton clouds must be absolutely identical with one another, as the exclusion principle seems to imply. Surely the physics cries out for some alternative picture of phenomena. It is the proton itself, not the space-time point, which behaves as a discrete physical entity and which has, at least to a considerable degree, some semblance of indivisibility. If a clear mathematical description can be found, according to which it is the space-time points that are "smeared out", rather than the particles, then I feel that we shall be closer to achieving a proper understanding of Nature.

It seems clear, however, that protons (or, indeed, any other known particles - at least, massive ones) are themselves somewhat composite objects and should not be used directly as the basic concepts with which to build up a mathematical formalism. What is required, perhaps, is something that lies rather between the concepts "particle" and "point". Indeed, a twistor has just such characteristics. It may be pictured in a geometrical way as something resembling a (massless) particle (although it is mathematically closer to being a kind of "square root" of a particle). Yet it is used in a way which makes it analogous, more, to the concept "position". Thus a quantum particle is described not by a single twistor or set of twistors but by a (holomorphic) function of twistors - analogously to the way that, in standard quantum theory, particles may be described by "wave functions" of position. (Alternatively the analogy might be, rather, with "momentum". This corresponds to using the complex conjugate twistor variables.) Of course the success, or otherwise, of any proposed new formalism can hardly be expected to rest on such quasi-philosophical considerations. Its inner consistency, agreement with observation - and even, perhaps, its intrinsic merit on aesthetic grounds - must be entered into in detail.

2. Twistor formalism

Consider a finite physical system in special relativity.

Its most important physical attributes are its 4-momentum p_a and its 6-angular momentum M^{ab} $(= -M^{ba})$, defined relative to some origin O . A twistor* Z^α is an object which can be described, in a suitable coordinate frame, by four complex numbers

$$(Z^0,\ Z^1,\ Z^2,\ Z^3) \tag{1}$$

where the complex conjugate twistor $\bar{Z}_\alpha$ of Z^α is described by

$$(\bar{Z}_0,\ \bar{Z}_1,\ \bar{Z}_2,\ \bar{Z}_3) = (\overline{Z^2},\ \overline{Z^3},\ \overline{Z^0},\ \overline{Z^1}) \tag{2}$$

(the symbol $\overline{Z^2}$ denoting the complex conjugate of the complex number Z^2 , etc.). For a massless particle the 4-momentum and angular momentum components, in a standard Minkowski frame, are to be related to those of Z^α by[2,3]

$$\begin{aligned} p^0 &= \frac{1}{\sqrt{2}}(Z^3\bar{Z}_1 + Z^2\bar{Z}_0)\\ p^1 &= \frac{1}{\sqrt{2}}(Z^3\bar{Z}_1 - Z^2\bar{Z}_0)\\ p^2 &= \frac{1}{\sqrt{2}}(-Z^2\bar{Z}_1 - Z^3\bar{Z}_0)\\ p^3 &= \frac{i}{\sqrt{2}}(Z^2\bar{Z}_1 - Z^3\bar{Z}_0) \end{aligned} \tag{3a}$$

---

* The abstract index notation[4] is being employed here (although this does not affect any of the explicit formulae). Greek indices refer to the four-dimensional twistor space, lower case Latin ones to the Minkowski space-time and capital Latin indices to two-dimensional spin-space, the complex conjugate space using primed indices. The Minkowski metric has signature (+---) and the standard Minkowski components of a vector are related to its standard spinor components via $(V^{00'},\ V^{01'},\ V^{10'},\ V^{11'}) = 2^{-\frac{1}{2}}(V^0+V^1,\ V^2+iV^3;\ V^2-iV^3,\ V^0-V^1)$. The skew ϵ-spinors raise and lower indices according to $\xi^A = \epsilon^{AB}\xi_B$, $\xi_B = \xi^A\epsilon_{AB}$. The sign of the skew e_{abcd} is fixed by $e_{0123} = 1$ in a standard frame. Round brackets around indices denote symmetrization.

$$
\begin{aligned}
M^{01} &= -M^{10} = \tfrac{1}{2}\,(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 - Z^2\bar{Z}_2 + Z^3\bar{Z}_3) \\
M^{02} &= -M^{20} = \tfrac{1}{2}\,(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 - Z^2\bar{Z}_3 - Z^3\bar{Z}_2) \\
M^{03} &= -M^{30} = \tfrac{1}{2}\,(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 - Z^2\bar{Z}_3 + Z^3\bar{Z}_2) \\
M^{12} &= -M^{21} = \tfrac{1}{2}\,(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 + Z^2\bar{Z}_3 - Z^3\bar{Z}_2) \qquad (3b) \\
M^{32} &= -M^{23} = \tfrac{1}{2}\,(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 + Z^2\bar{Z}_2 - Z^3\bar{Z}_3) \\
M^{13} &= -M^{31} = \tfrac{1}{2}\,(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 + Z^2\bar{Z}_3 + Z^3\bar{Z}_2) \\
& M^{00} = M^{11} = M^{22} = M^{33} = 0 \,.
\end{aligned}
$$

Using the 2-spinor formalism we can express this as

$$P_{AA'} = \bar{\pi}_A \pi_{A'} \,, \qquad (4)$$

$$M^{AA'BB'} = i\omega^{(A}\bar{\pi}^{B)}\epsilon^{A'B'} - i\epsilon^{AB}\bar{\omega}^{(A'}\pi^{B')} \,, \qquad (5)$$

where the components of the 2-spinors ω^A, $\pi_{A'}$ in the standard frame are

$$Z^0 = \omega^0 \,, \quad Z^1 = \omega^1 \,, \quad Z^2 = \pi_{0'} \,, \quad Z^3 = \pi_{1'} \qquad (6)$$

which (in a frame-independent manner, with respect to the origin O) may be written as:

$$Z^\alpha \overset{O}{\leftrightarrow} (\omega^A, \pi_{A'}) \,. \qquad (7)$$

The complex conjugate twistor is then given by

$$\bar{Z}_\alpha \overset{O}{\leftrightarrow} (\bar{\pi}_A, \bar{\omega}^{A'}) \,. \qquad (8)$$

The representation with respect to a new (real) origin Q , whose position vector with respect to O is x^a (= $x^{AA'}$) is

$$Z^\alpha \overset{Q}{\leftrightarrow} (\omega^A - ix^{AA'}\pi_{A'}, \pi_{A'}) \qquad (9)$$

$$\bar{Z}_\alpha \overset{Q}{\leftrightarrow} (\bar{\pi}_A, \bar{\omega}^{A} + ix^{AA'}\bar{\pi}_A) \,, \qquad (10)$$

this being consistent with the fact that the 4-momentum and angular-momentum with respect to Q are given, respectively, by

$$p_a \ , \quad M^{ab} - x^a p^b + x^b p^a \ . \tag{11}$$

The Pauli-Lubanski spin-vector

$$S_a = \tfrac{1}{2} e_{abcd} \, p^b M^{cd} \tag{12}$$

is, like p_a, independent of the choice of origin. For a massless particle we have

$$S_a = s p_a \tag{13}$$

where s is the helicity. Equation (13) is automatically satisfied by (3) (i.e. by (4) and (5)) with

$$s = \tfrac{1}{2}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha) \ . \tag{14}$$

In obtaining (14), it has not been assumed that Z^α and $\bar{Z}_\alpha$ commute but that *either* all components commute (classical twistor variables) *or* that the twistor ("first-quantization") commutation rules

$$[Z^\alpha, Z^\beta] = 0 \ , \quad [\bar{Z}_\alpha, \bar{Z}_\beta] = 0 \ , \quad [Z^\alpha, \bar{Z}_\beta] = \delta^\alpha_\beta \tag{15}$$

hold (taking $\hbar = 1$). The normal commutation rules for p_a and M^{ab}, which give rise to the Lie algebra of the Poincaré group are, in fact, automatic consequences of (15).

Reverting to the classical case of commuting variables, we may obtain a geometrical picture of a twistor Z^α, in the case $s = 0$, by asking for the locus Z of points Q with respect to which the description of Z^α has the form $(0, \pi_{A'})$, i.e. for which

$$\omega^A = i x^{AA'} \pi_{A'} \ . \tag{16}$$

These points are the ones about which the angular momentum vanishes, so Z describes the history of the mass-centre of the

the particle and is, in fact, a null straight line in the direction of p_a (assuming $\pi_{A'} \neq 0$). The situation when $s \neq 0$ is more complicated[1,3] and will not be entered into here.

If $s = 0$, we refer to Z^α as a null twistor. If Z^α and Y^α are two null twistors then the necessary and sufficient condition for the corresponding null lines Z and Y to intersect (possibly at conformal infinity[1,3,4]) is

$$Z^\alpha \bar{Y}_\alpha = 0 \qquad (17)$$

in which case the system of null lines X through their intersection point P is given by the linear set of twistors

$$X^\alpha = \lambda Z^\alpha + \mu Y^\alpha , \qquad (18)$$

where λ and μ are complex numbers, not both zero.

Thus points of Minkowski space are represented (provided our twistors remain classical) by the 2-dimensional complex vector subspaces of twistor space which are given by (18), Z^α and Y^α being both null and subject to (17). The space-time point P is, in this way, represented in terms of the system of twistors "through P". We do not have to introduce points as primary objects; the twistors themselves are taken to be primary and the points defined in terms of such systems of twistors. In the classical domain this gives us merely an alternative mathematical framework for the description of space-time geometry. But with twistor quantum operators we now have the possibility of treating points as "smeared out" objects.

Under the action of the conformal group $C(1,3)$ applied to the space-time, the classical twistor components $Z^0, \ldots, Z^3$ are transformed[*] by the pseudo-unitary matrices of $SU(2,2)$ where the invariant Hermitian form is $Z^\alpha \bar{Z}_\alpha$ $(= 2s)$. The explicit action of the Poincaré subgroup is implicit in the origin transformation behaviour (9), (10), together with the SL(2,C) transformations on the spinor components which arise when Lorentz rotations are applied.

To describe a massive particle or system, a set of two or more twistors

$$\underset{1}{Z}^{\alpha},\ \underset{2}{Z}^{\alpha},\ \dots,\ \underset{r}{Z}^{\alpha} \tag{19}$$

is required. The momentum and angular momentum are now defined by expressions like (3) but where the right-hand sides are replaced by sums of r similar expressions, involving each one of the twistors (19) in turn. Using the notation of (4) and (5), this may be written:

$$\begin{aligned} p_{AA'} &= \underset{1}{\bar{\pi}}_A \underset{1}{\pi}_{A'} + \dots + \underset{r}{\bar{\pi}}_A \underset{r}{\pi}_{A'}\,, \\ M^{AA'BB'} &= i\underset{1}{\omega}^{(A}\underset{1}{\bar{\pi}}^{B)}\varepsilon^{A'B'} - i\varepsilon^{AB}\underset{1}{\bar{\omega}}^{(A'}\underset{1}{\pi}^{B')} + \dots + \\ &\quad + i\underset{r}{\omega}^{(A}\underset{r}{\bar{\pi}}^{B)}\varepsilon^{A'B'} - i\varepsilon^{AB}\underset{r}{\bar{\omega}}^{(A'}\underset{r}{\pi}^{B')}\,. \end{aligned} \tag{20}$$

In fact $r = 2$ is sufficient for defining the momentum and angular momentum of a general massive system. However, larger values of r seem to be required in order to express internal degrees of freedom.

The notation is helped by the introduction of the angular momentum twistor:

$$A^{\alpha\beta} \leftrightarrow \begin{pmatrix} -2i\mu^{AB} & p^{A}{}_{B'} \\ p_{A'}{}^{B} & 0 \end{pmatrix}\,, \tag{21}$$

where

$$M^{AA'BB'} = \mu^{AB}\varepsilon^{A'B'} + \varepsilon^{AB}\bar{\mu}^{A'B'} . \qquad (22)$$

Then (20) may be expressed

$$A^{\alpha\beta} = 2\underset{1}{Z}^{(\alpha}I^{\beta)\gamma}\underset{1}{\bar{Z}}_{\gamma} + \dots + 2\underset{r}{Z}^{(\alpha}I^{\beta)\gamma}\underset{r}{\bar{Z}}_{\gamma} \qquad (23)$$

the skew-symmetrical infinity twistors $I_{\alpha\beta}$, $I^{\alpha\beta}$ being defined by

$$I_{\alpha\beta} \leftrightarrow \begin{pmatrix} 0 & 0 \\ 0 & \varepsilon^{A'B'} \end{pmatrix} , \quad I^{\alpha\beta} \leftrightarrow \begin{pmatrix} \varepsilon^{AB} & 0 \\ 0 & 0 \end{pmatrix} \qquad (24)$$

(The presence of an infinity twistor in an expression such as (23) signifies the fact[1,2] that conformal invariance has been reduced to Poincaré invariance.) The angular momentum twistor possesses the symmetries

$$A^{\alpha\beta} = A^{\beta\alpha} , \quad A^{\alpha\gamma}I_{\beta\gamma} = \bar{A}_{\beta\gamma}I^{\alpha\gamma} . \qquad (25)$$

For each value of r there is an "internal" symmetry group arising from the linear transformations of $\underset{1}{Z}^{\alpha}, \dots, \underset{r}{Z}^{\alpha}, \underset{1}{\bar{Z}}_{\alpha}, \dots, \underset{r}{\bar{Z}}_{\alpha}$ which preserve (23). When $r = 1$ this is the $U(1)$ of phase transformations

$$Z^{\alpha} \to e^{i\theta}Z^{\alpha} , \quad \bar{Z}_{\alpha} \to e^{-i\theta}\bar{Z}_{\alpha} \qquad (26)$$

(θ real), the infinitesimal generator of which may be taken to be the helicity operator s of (14) (adopting (15)). When $r = 2$ we obtain (denoting the twistors by Z^{α}, X^{α})

$$\begin{pmatrix} Z^{\alpha} \\ X^{\alpha} \\ \bar{Z}_{\alpha} \\ \bar{X}_{\alpha} \end{pmatrix} \mapsto \begin{pmatrix} \begin{bmatrix} \rho & \sigma \\ \tau & \kappa \end{bmatrix}\delta^{\alpha}_{\beta} & \begin{bmatrix} 0 & \lambda \\ -\lambda & 0 \end{bmatrix}\begin{bmatrix} \rho & \sigma \\ \bar{\tau} & \bar{\kappa} \end{bmatrix}I^{\alpha\beta} \\ \begin{bmatrix} 0 & \lambda \\ -\bar{\lambda} & 0 \end{bmatrix}\begin{bmatrix} \rho & \sigma \\ \tau & \kappa \end{bmatrix}I_{\alpha\beta} & \begin{bmatrix} \rho & \sigma \\ \bar{\tau} & \bar{\kappa} \end{bmatrix}\delta^{\beta}_{\alpha} \end{pmatrix} \begin{pmatrix} Z^{\beta} \\ X^{\beta} \\ \bar{Z}_{\beta} \\ \bar{X}_{\beta} \end{pmatrix} \qquad (27)$$

where the matrix of ρ, σ, τ, κ is unitary, the group (27) being isomorphic with a double covering of O(3) × E(2) (where E(2) is the group of orientation preserving Euclidean motions of the plane). The infinitesimal generators may be taken as

$$\begin{aligned} j &= \tfrac{1}{2}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha - X^\alpha \bar{X}_\alpha - \bar{X}_\alpha X^\alpha)\ , \\ k &= \tfrac{1}{2}(X^\alpha \bar{Z}_\alpha + Z^\alpha \bar{X}_\alpha)\ , \quad \ell = \tfrac{i}{2}(Z^\alpha \bar{X}_\alpha - X^\alpha \bar{Z}_\alpha) \end{aligned} \qquad (28)$$

for the SU(2) part and

$$\begin{aligned} \mu &= I_{\alpha\beta} Z^\alpha X^\beta\ , \quad \bar{\mu} = I^{\alpha\beta} \bar{Z}_\alpha \bar{X}_\beta \\ q &= \tfrac{1}{2}(Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha + X^\alpha \bar{X}_\alpha + \bar{X}_\alpha X^\alpha) \end{aligned} \qquad (29)$$

For the $\tilde{E}(2)$ part. In each case the commutation rules for the twistor variables are as given in (15), where the only non-vanishing commutators occur between each twistor and its complex conjugate. We have

$$m^2 = 2\mu\bar{\mu}\ , \quad -S_a S^a m^{-2} = j(j+1) + (k-i\ell)(k+i\ell) \qquad (30)$$

for the Casimir operators describing squared mass[2] and spin[15] respectively. The situation for three or more twistors proceeds in a similar fashion but is more complicated[10]. The matrices analogous to those of (27), in the general case, involve an (r × r) unitary and an (r × r) complex skew-symmetric matrix.

To describe the state of a particle we require a holomorphic function of all the variables occurring, where for each twistor we select either the twistor itself or its complex conjugate as argument:

$$F(\underset{1}{Z^\alpha}, \ldots, \underset{u}{Z^\alpha}, \underset{u+1}{\bar{Z}}{}_\alpha, \ldots, \underset{r}{\bar{Z}}{}_\alpha)\ . \qquad (31)$$

To fix the quantum numbers we may require F to be in an eigenstate of an appropriate selection of operators such as a commuting set from those occurring in (28), (29), (30). We may then construct a space-time field from F by means of a contour integral. The value of the field at the point Q, with (possibly complex) position vector x^a, is obtained by choosing $\underset{1}{Z}^{\alpha}, \ldots, \underset{r}{\bar{Z}}_{\alpha}$ to be twistors "through Q" so that

$$\begin{aligned} \underset{p}{\omega}^{A} &= i x^{AA'} \underset{p}{\pi}_{A'} \quad (p = 1, \ldots, u) \\ \underset{p}{\bar{\omega}}^{A'} &= -i x^{AA'} \underset{p}{\bar{\pi}}_{A} \quad (p = u+1, \ldots, r) \end{aligned} \tag{32}$$

in accordance with (16). Then using some suitable 2r-dimensional contour surrounding singularities of F, we perform

$$\phi_{A' \ldots C' D \ldots F G' \ldots K \ldots M'} = \frac{1}{(2\pi i)^{2r}} \oint \underset{1}{\pi}_{A'} \ldots \underset{1}{\pi}_{C'} \frac{\partial}{\partial \underset{1}{\omega}^{D}} \ldots \frac{\partial}{\partial \underset{1}{\omega}^{F}} \underset{2}{\pi}_{G'} \ldots$$

$$\ldots \underset{v}{\bar{\pi}}_{K} \ldots \frac{\partial}{\partial \bar{\omega}^{M'}} F \, \delta Z , \tag{33}$$

where

$$\delta Z = d\underset{1}{\pi}_{A'} \wedge d\underset{1}{\pi}^{A'} \wedge d\underset{2}{\pi}_{B'} \wedge d\underset{2}{\pi}^{B'} \wedge \ldots \wedge d\underset{r}{\bar{\pi}}_{E} \wedge d\underset{r}{\bar{\pi}}^{E} . \tag{34}$$

The precise choice of how many $\pi_{A'}$ terms and $\partial/\partial\omega^A$ terms appear in (33) is up to our discretion. However, fixing the appropriate quantum numbers for F may have the implication that F has a definite homegeneity degree in each of its twistor arguments. In this case only certain of the various expressions (33) can give non-zero answers, namely those in which the total homegeneity degree in the integrand (including the differential δZ) is zero in each twistor variable. (Thus we count +1 for each $\pi_{A'}$, -1 for each $\partial/\partial\omega^A$ and +2 for $d\pi_{A'} \wedge d\pi^{A'}$. This total, added to the homogeneity degree of F in the corresponding twistor variable, has to be zero, for a non-zero resulting field.) When F is homogeneous, the 2n-dimensional integral (33) can be reduced to an r-dimensional one where each $d\pi_{A'} \wedge d\pi^{A'}$ is replaced by $\pi_{A'}d\pi^{A'}$.

Various special cases of (33) have been given earlier[2]. (The incorporation of the $\partial/\partial\omega^A$ terms follows a suggestion made by L. P. Hughston.) If $r = 1$ and if F is in an eigenstate of s, as given in (14), then the eigenvalue is one of the values $\ldots, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, \ldots,$ giving the helicity; and F has integral homogeneity. The field given by (33) then turns out to be automatically a solution of the massless free field equations for this helicity (assuming that it has positive frequency, otherwise the sign of the helicity is reversed). The procedure enables us to generate arbitrary (analytic) solutions of the wave equation, Weyl neutrino equation, Maxwell free-field equations and linearized Einstein equations starting from an arbitrary homogeneous holomorphic fuction of a twistor. If $r = 2$ and F is in an eigenstate of j, q and m^2 then we can generate solutions of the Dirac ("higher spin") irreducible equation[16] for spin j if we use a minimal number of π_A's and $\partial/\partial\omega^A$'s and for various spins $> j$ otherwise. If we further fix F to be in an eigenstate of $(k-i\ell)(k+i\ell)$ then we get only one such field (or its n^{th} derivative) - the "minimal" one if this eigenvalue is zero.

Since (33) is a contour integral there is the "gauge" freedom that if any holomorphic function is added to F which is non-singular "inside" the chosen contour, then the result of the integral is unchanged. This, in effect, tells us that any physically measurable quantity calculated from F must be obtainable by means of a contour integration. In particular, the functions F may be employed as the in- and out-states for calculations of scattering amplitudes, where these calculations are to be carried out by performing certain many-dimensional contour integrations. The beginnings of a general scheme for doing this are indicated elsewhere[2,3].

References

1. R. Penrose, J. Math. Phys. 8 (1967) 345.
2. R. Penrose & M.A.H. MacCallum, Physics Reports 6C (1973) 242.
3. R. Penrose, in Quantum Gravity (Eds. C.J. Isham, R. Penrose & D.W. Sciama, Clarendon Press, Oxford, 1975).
4. R. Penrose, Int. J. Theor. Phys. 1 (1968) 61.
5. R. Penrose, in Magic without Magic, John Archibald Wheeler (Ed. J.R. Klauder; W.H. Freeman & Co., San Francisco 1972).
6. R. Penrose, in Quantum Theory and Beyond (Ed. Ted Bastin; Cambridge University Press, 1971).
7. E. Cartan, Ann. Sci. Ecole Normale Superieure 31 (1914) 263.
8. R. Penrose, in Group Theory in Non-Linear Problems (Ed. A.O. Barut; D. Reidel 1974).
9. J.A. McLennan, Jr., Nuovo Cim. 10 (1956) 1360.
10. Z. Perjés, Twistor Variables of Relativistic Mechanics (Preprint, 1974).
11. J.A. Wheeler, Geometrodynamics (Academic Press, New York 1962).
12. F.J.M. Farley, Nuovo Cim. 1, Numero Speciale (1969) 59.
13. B. Richter, Report at XVII Int. Conf. on High Energy Physics, London SLAC-PUB-1478 (1974).
14. R. Penrose, in Battelle Rencontres 1967 (Eds. C.M. DeWitt & J.A. Wheeler; Benjamin, New York 1968).

15. P. Tod (to appear).

16. P.A.M. Dirac, Proc. Roy. Soc. (Lond.) A155 (1936) 447.

67

THE NON-LINEAR GRAVITON

Gravity Research Foundation Essay Prize, 1975

This was the third time I entered an essay for the Gravity Research Foundation Prize. I wrote this one on my own, however, and it rather surprisingly won the first prize! (very much more recently I put in for a fourth time and all I got was an 'honorable mention'—which, on reflection, I would agree was about appropriate.) This current essay was later published (Chapter 71) almost without change. Chapter 70 is the serious write-up.

THE NON-LINEAR GRAVITON

BY

ROGER PENROSE

Mathematical Institute

University of Oxford

Oxford, England

First Award Winning Essay
of the
Gravity Research Foundation
58 Middle Street
Gloucester, Mass. 01930

1975

Summary

A new approach to quantized gravitational theory is suggested. It is argued by analogy with Maxwell theory - and also from a principle that (physical) gravitons should carry space-time curvature - that a free graviton should be describable by a complex solution of Einstein's vacuum equations. For a left-handed graviton (pure helicity state $-2\hbar$) we require a solution which is of a particular kind, called right-flat, and which is of positive frequency. A construction is given for obtaining all such solutions, in terms of general curved (top halves of) twistor spaces.

Despite impressive progress, recently(1), towards the intended goal of a satisfactory quantum theory of gravity, there remain fundamental problems whose solutions do not appear to be yet in sight. In particular, the standard perturbation techniques seem to lead to a quantized Einstein theory which is not renormalizable(1), and it has consequently been argued that Einstein's equations should perhaps be replaced by something more compatible with conventional quantum field theory. There is also the alternative possibility, which has occasionally been aired, that some of the basic principles of quantum mechanics may need to be called into question(2) if we are ever to arrive at a completely satisfactory quantum theory of gravity. It is this second alternative that I wish to explore here; and it is the principle of linear superposition (at least in its standard form) that I shall be calling into question.

Let us first ask: what is the most primitive physical requirement of a quantum theory? As the name implies, what we expect is that it should provide us with quanta - the particles associated with whatever field or fields we are attempting to describe. Thus, a quantum theory of gravity should provide gravitons.

Furthermore, since we are trying to do physics, rather than mathematical exercises, we require that a really physical concept of graviton actually does exist. For the purpose of this essay I am assuming this to be the case. I hope that the somewhat remarkable mathematical result that I shall give lends some independent theoretical support for this supposition.

I shall proceed by analogy with the electromagnetic field, at first. How, indeed, do we provide a mathematical description of a single free photon? I require a description which is both relativistically invariant and gauge invariant. The wave function can then be given as a solution F_{ab} $(= -F_{ba})$ of the free Maxwell equations

$$\nabla_{[a}F_{bc]} = 0 , \qquad \nabla^{a}F_{ab} = 0 . \tag{1}$$

However, there are two essential differences between classical solutions of the equations (1) and solutions giving wave functions for free photons. In the first place, classical solutions are real while wave functions are complex; in the second (and related to the first), wave functions should be composed only of positive frequencies*. There is also the question of choosing a normalization for the wave function, but I shall not consider that here, as it is not strictly necessary at this stage. But one can go further than this. The description so far is (in our world, where parity conservation is not a general law of nature) a reducible one. We may regard it as describing two different particles simultaneously, namely the left-handed and the right-handed photon. To examine this, consider the decomposition of F_{ab} into its self-dual and anti-self-dual parts:

---

* There are, of course, many alternative equivalent ways of describing a free photon. The choice I am making here seems to me to be the most natural - and also the only one which fits in with the further ideas I am suggesting.

$$F_{ab} = F^{(+)}_{ab} + F^{(-)}_{ab}$$

where

$$F^{(\pm)}_{ab} = \pm i\, e_{abcd}\, F^{(\pm)cd}$$

e_{abcd} being the alternating tensor. If F_{ab} is real, $F^{(+)}_{ab}$ and $F^{(-)}_{ab}$ are complex conjugates of one another, but for a wave function, they are simply independent complex quantities, each of positive frequency (and therefore necessarily not complex conjugates of one another). In fact $F^{(+)}_{ab}$ describes a left-handed photon (helicity $-\hbar$) and $F^{(-)}_{ab}$ a right-handed photon (helicity $+\hbar$) .

A spinor description can also be used. Put

$$F_{ab} = \phi_{AB}\, \varepsilon_{A'B'} + \varepsilon_{AB}\, \tilde{\phi}_{A'B'} \qquad (2)$$

(the abstract index notation being employed, for convenience), where

$$F^{(+)}_{ab} = \phi_{AB}\, \varepsilon_{A'B'} \qquad F^{(-)}_{ab} = \varepsilon_{AB}\, \tilde{\phi}_{A'B'} \ .$$

Then for a classical (real) field we have $\phi_{A'B'} = \bar{\phi}_{A'B'}$ whereas for a photon wave function ϕ_{AB} and $\tilde{\phi}_{A'B'}$ are independent (symmetric) spinors describing the left-handed and right-handed components of the photon(4). The Maxwell equations (1) become

$$\nabla^{AA'}\, \phi_{AB} = 0\ , \qquad \nabla^{AA'}\, \tilde{\phi}_{A'B'} = 0\ . \qquad (3)$$

To describe a pure $-\hbar$ helicity state we must take $\tilde{\phi}_{A'B'} = 0$; for a pure $+\hbar$ helicity state we take $\phi_{AB} = 0$. The requirement that ϕ_{AB} and $\tilde{\phi}_{A'B'}$ have positive frequency may be stated in the following geometrical way: it is possible to extend the domain of definition of each field quantity holomorphically to points of complexified Minkowski space $\mathbb{C}\mathbb{M}$, whose position vectors $x^a + i\, y^a$ have imaginary part y^a lying on or within the past null cone.(3) This portion of $\mathbb{C}\mathbb{M}$ will be denoted: $\mathbb{C}\mathbb{M}^+$

Let us now turn to the case of gravity. There is an obvious analogue to the above description, where two four-index symmetric spinors ϕ_{ABCD}, $\tilde{\phi}_{A'B'C'D'}$ are employed in a Minkowski background space to provide a (complex) linearized curvature tensor

$$K_{abcd} = \phi_{ABCD}\, \varepsilon_{A'B'}\, \varepsilon_{C'D'} + \varepsilon_{AB}\, \varepsilon_{CD}\, \tilde{\phi}_{A'B'C'D'}$$

The discussion is exactly as above. A common view of a graviton has, indeed, been essentially this, namely something which refers to weak-field gravitational perturbations of a space-time (normally flat). However, it is at this point that I wish to take issue with the standard view. I cannot (now) believe that a physical graviton should be described by linear gravitational theory. My reasoning is as follows. If we start with flat space, and then add one (linearized) "graviton", the space remains flat. The null cones have not shifted. We add a second such "graviton" and a third and a fourth, and the space is still flat with null cones still locked in their original Minkowskian positions. With such a perturbative viewpoint it is only after an infinite number of "gravitons" have been added that the space can become curved. The situation may be compared with a power series expansion. For any finite number of terms the function

$$1 + \frac{1}{z} + \frac{1}{z^2} + \ldots + \frac{1}{z^n}$$

has a pole singularity locked at the position $z = 0$. But the sum to infinity

$$1 + \frac{1}{z} + \frac{1}{z^2} + \frac{1}{z^3} + \ldots \quad = \quad \frac{z}{z-1}$$

has its pole shifted to the position $z = 1$. The usual perturbative expansions of graviton propagators have a similar behaviour, with poles corresponding to the Minkowskian null cones right up until the sum to infinity has been performed. Surely, if it has meaning to talk about physical gravitons, it should have meaning to talk about finite numbers of such gravitons. And if Einstein's theory is physically appropriate, then surely each graviton itself carries its measure of curvature. I cannot believe that curvature emerges only after an infinite sum has been performed.

So rather than working with linear theory, let us attempt to apply the electromagnetic analogy instead to the full non-linear Einstein theory. Indeed, a decomposition of the Riemann curvature of a vacuum space-time can be expressed in a form analogous to (2):

$$R_{abcd} = \Psi_{ABCD}\,\varepsilon_{A'B'}\,\varepsilon_{C'D'} + \varepsilon_{AB}\,\varepsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'} \;;$$

where the two terms on the right are the self-dual and anti-self-dual parts of the vacuum curvature, respectively. For a classical (real) space-time, $\tilde{\Psi}_{A'B'C'D'}$ must be the complex conjugate of Ψ_{ABCD}. But to describe a (non-linear) graviton we must expect, by analogy with the electromagnetic case, that Ψ_{ABCD} and $\tilde{\Psi}_{A'B'C'D'}$ will become independent quantities each, in some appropriate sense, of positive frequency. The curvature R_{abcd} becomes complex, so we are dealing with some form of complex Riemannian space. The metric g_{ab} must, indeed, be complex symmetric (holomorphic, not Kähler). The complex Ricci tensor must vanish. The complex vacuum Bianchi identity becomes

$$\nabla^{AA'}\,\Psi_{ABCD} = 0\ , \qquad \nabla^{AA'}\,\tilde{\Psi}_{A'B'C'D'} = 0$$

in close analogy with (3) .

But if such a situation describes a graviton it would generally be a graviton of mixed helicities. For a pure graviton in a helicity state we would apparently require $\tilde{\Psi}_{A'B'C'D'} = 0$ (helicity $-2\hbar$) or else $\Psi_{ABCD} = 0$ (helicity $+2\hbar$). I shall refer to a complex vacuum space-time for which $\tilde{\Psi}_{A'B'C'D'} = 0$ as right-flat; and one for which $\Psi_{ABCD} = 0$ as left-flat. Thus, I am proposing that a single left-handed graviton is described by a positive-frequency right-flat (left-curved) complex space-time; and a right-handed graviton by a left-flat space-time*.

Such a description may seem somewhat mysterious (and "positive-frequency" has yet to be clearly defined in this context). But it is, to my mind, a very remarkable fact that the general such complex space-time can be constructed automatically by way of twistor theory(4),(1)! There are already various reasons for believing that twistor theory may be able to provide some new insights into the workings of nature; and here we have an apparently independent role for this theory. It will not be possible to go into any detail, but the essential mathematical result can be briefly described.

Let us first recall(4) how complex (compactified) Minkowski space $\mathbb{CM}$ arises from standard (flat) twistor space $\mathbb{T}$. The space $\mathbb{T}$ is a complex four-dimensional vector space endowed with a Hermitian form Σ of signature $(++--)$. The parts of $\mathbb{T}$ at which Σ is positive, negative, or zero are labelled $\mathbb{T}^+$, $\mathbb{T}^-$, $\mathbb{N}$, respectively. The complex two-dimensional linear subspaces of $\mathbb{T}$ correspond to the points of $\mathbb{C}\mathbb{M}$; those lying in $\mathbb{N}$ define the points of real (compactified) Minkowski space $\mathbb{M}$; those lying in $\mathbb{T}^+ \cup \mathbb{N}$ define the points of $\mathbb{C}\mathbb{M}^+$. Often it is convenient to consider the projective twistor space $\mathbb{P}\mathbb{T}$ (where the prefix $\mathbb{P}$ means "projective"). The points of $\mathbb{C}\mathbb{M}$ are then represented in $\mathbb{P}\mathbb{T}$ by complex projective lines (compact holomorphic curves having topology S^2); the points of $\mathbb{M}$ by projective lines in $\mathbb{P}\mathbb{N}$ and those of $\mathbb{C}\mathbb{M}^+$ by projective lines in $\mathbb{P}\mathbb{T}^+ \cup \mathbb{P}\mathbb{N}$. The space $\mathbb{T}$ is also endowed with a (Poincaré invariant) 1-form and 2-form:

---

* The concepts of right-flat and left-flat space-times arose historically from E. T. Newman's construction of the spaces(5) referred to as "Heavens". Though much of this essay has its origins in Newman's construction, I shall not have need to refer to it explicitly here.

$$\underline{\iota} = I_{\alpha\beta} Z^{\alpha} dZ^{\beta}, \quad \underline{\tau} = d\underline{\iota} = I_{\alpha\beta} dZ^{\alpha} \wedge dZ^{\beta}$$

(the form $\underline{\tau}$ being degenerate, of rank 2); and with (a conformally invariant) 3-form and 4-form:

$$\underline{\rho} = \varepsilon_{\alpha\beta\gamma\delta} Z^{\alpha} dZ^{\beta} \wedge dZ^{\gamma} \wedge dZ^{\delta}$$

$$\underline{\omega} = d\underline{\rho} = \varepsilon_{\alpha\beta\gamma\delta}\, dZ^{\alpha} \wedge dZ^{\beta} \wedge dZ^{\gamma} \wedge dZ^{\delta} \quad ,$$

in standard twistor notation(1).

To construct a left-handed non-linear graviton we proceed as follows. We consider the bounded portion $\mathbb{P}\mathbb{T}^+ \cup \mathbb{P}\mathbb{N}$ of projective twistor space, deleting the part $\mathbb{P}\mathbb{T}^-$. We then perturb its complex structure in an arbitrary fashion to obtain a new space $\mathscr{D}$. By appealing to a theorm of Kodaira(6)(and employing some observations due to M. F. Atiyah) we can infer the existence of a complex four-parameter family of compact holomorphic curves in $\mathscr{D}$, which are the analogues of the projective lines in $\mathbb{P}\mathbb{T}^+$.

Each such curve now defines a point in an abstract complex four-manifold $\mathcal{M}$. The space $\mathcal{M}$ naturally acquires a (complex) conformal structure, where null separation in $\mathcal{M}$ corresponds to intersection of the corresponding curves in $\mathscr{P}$. This conformal structure can be shown to be Riemannian (by general theory) with a right-flat conformal curvature. We now require that the forms $\underline{\iota}$, $\underline{\tau}$, $\underline{\rho}$, $\underline{\omega}$ have been introduced, with the same stated properties as for flat twistors, but otherwise arbitrary, We apparently need solve no equations to achieve all this. But, remarkably, the resulting complex four-space $\mathcal{M}$ acquires the structure of a general right-flat solution of the Einstein vacuum equations!

To ensure that $\mathcal{M}$ is, in some appropriate sense, of positive frequency, we require that $\mathscr{P}$ is sufficiently extended that it is validly analogous to $\mathbb{PT}^+ \cup \mathbb{PN}$. I cannot explain the details here, but the requirement seems to be that the boundary of $\mathscr{P}$ is ruled appropriately by a real three-parameter family of holomorphic curves. Again, it seems that no equations need be solved to achieve this. Details will be published elsewhere.

There is clearly much more that needs to be done for this line of approach to quantum gravity to be made the basis of a physical theory. We require a definition of scalar product between these one-particle states. And how are we to build up a many-graviton theory from this one-graviton approach*? Further work is in progress.

My thanks are especially due to E. T. Newman and M. F. Atiyah for supplying vital ideas.

---

* A modified form of Geroch's proposal(7) seems worth exploring here. I am grateful to A. Ashtekar for suggesting this.

(1) C. J. Isham, R. Penrose and D. W. Sciama (Eds.), *Quantum Gravity: An Oxford Symposium* (Oxford Univ. Press, 1975).

(2) B. Mielnik, *Commun. Math. Phys*. **37** (1974) 221.

(3) R. F. Streater and A. S. Wightman, *PCT, Spin and Statistics, and All That* (W. A. Benjamin, Inc., New York, 1964).

(4) R. Penrose and M.A.H. MacCallum, *Physics Reports* **6C** (1973) 242.

(5) E. T. Newman, report to Tel Aviv Conference 1974.

(6) K. Kodaira, *Amer. J. Math*. **85** (1963) 79.

(7) R. P. Geroch, *Comm. Math. Phys*. **26** (1972) 271.

68

CLASSICAL GRAVITY: A STATUS REPORT

Trieste Summer School Proceedings. Eds R. Ruffini *et al.* ICTP, 1975

CLASSICAL GRAVITY: A STATUS REPORT

R. Penrose
Mathematical Institute, University of Oxford, Oxford, England.

Perhaps the most outstanding problem of classical general relativity is to decide whether the "cosmic censorship principle" holds (1, 2). This principle asserts that the space-time singularities resulting from gravitational collapse must be invisible to observers at large distances, leading to the now standard picture of a black hole. Work by Teokolsky and others (3-6) points strongly to the stability of (classical) Kerr black holes, these previously having been virtually established (through the work of Israel (7), Carter (8), Hawking (9) and others (10)) as the only possible stationary black-hole configurations. Although special models can be constructed which violate cosmic censorship (11), these cannot be regarded as totally plausible, physically. Likewise, all direct attempts at a contradiction with the black hole picture have so far failed (2, 12). Computer calculations by De Witt and his colleagues on the congealing of two black holes should shed some light on this problem, but they are not yet completed. The extreme case of a relativistic head-on collision has been studied by geometric methods, and an upper limit on the energy of emitted gravitational waves of one half the initial rest energy has been obtained (12).

While a satisfactory definition of energy and momentum for an isolated system has been known for some time from the works of Bondi (13) and Sachs (14), with the important implication that the energy of gravitational waves is necessarily positive, the status of angular momentum has remained somewhat problematical (15, 16). The difficulty arises from the fact that the asymptotic symmetries of an asymptotically flat space-time differ from those of Minkowski space and give rise, not to the Poincaré group, but to an infinite-parameter extension of it, known as the Bondi-Metzner-Sachs group (13, 14, 17-19). However, an important new advance due to Newman (20) has established the (complex) Riemannian nature of a related space $\mathscr{H}$, which may be regarded as the carrier space of asymptotic massless fields (21). It now turns out that $\mathscr{H}$ satisfied Einstein's vacuum equations, whether or not the original space does, and has some other simplifying properties. It seems likely that significant new developments in the angular momentum problem may be close at hand.

References

1. R. Penrose, Rev. Nuovo Cim. (Num. Spec.) **1** (1969) 252.
2. Penrose, Ann. N. Y. Acad. Sci. **224** (1973) 125.
3. S. Teukolsky, Astrophys. J. **185** (1973) 635.
4. W. H. Press & S. Teukolsky, Astrophys. J. **185** (1973) 649.
5. J. B. Hartle & D. C. Wilkins, "Analytic Properties of the Teukolsky Equation" (preprint 1974).
6. J. M. Stewart, "On the Stability of Kerr's Spacetime" (preprint 1974).
7. W. Israel, Phys. Rev. **164** (1967) 1776.
8. B. Carter, Phys. Rev. Lett. **78** (1971) 331.
9. S. W. Hawking, Commun. Math. Phys. **25** (1972) 152.
10 H. Müller zum Hagan, D. C. Robinson & H.-J. Seifert, Gen. Rel. & Grav. **4** (1972) 1, 53.
11. P. Yodzis, H.-J. Seifert & H. Müller zum Hagen, Comm. Math. Phys. **34** (1973) 135; **37** (1974) 29.
12. R. Penrose, "Black Holes in Relativistic Collision" (to appear).
13. H. Bondi, M. G. J. van der Burg & A. W. K. Metzner, Proc. Roy. Soc. (Lond) A **269** (1962) 21.
14. R. K. Sachs, Proc. Roy. Soc. (Lond) A **270** (1962) 103.
15. R. W. Lind, J. Messmer & E. T. Newman, J. Math. Phys. **13** (1972) 1884.
16. B. D. Bramson, Proc. Roy. Soc. (Lond) A **341** (1975), 463.
17. R. K. Sachs, Phys. Rev. **128** (1962) 2851.
18. P. J. McCarthy, Proc. Roy. Soc. (Lond) A **330** (1972) 517; **333** (1973) 317.
19. R. Penrose in "Group Theory in Non-Linear Problems" (Ed. A. O. Barut; Reidel, Dordrecht 1974).
20. E. T. Newman, (report to Tel Aviv conference, 1974).
21. B. Aronson, R. W. Lind, J. Messmer & E. T. Newman, J. Math. Phys. **12** (1971) 2462.

69

TWISTOR THEORY: ITS AIMS AND ACHIEVEMENTS

Quantum gravity, an Oxford symposium. Eds C. J. Isham,
R. Penrose, and D. W. Sciama, pp. 268–407. Oxford University
Press, 1975

This survey of the status of twistor theory introduces some new ingredients. It begins with some early motivational links with spin-network theory (Chapter 31) that had not been expressed before. Among the other novel ingredients is a description of the procedure that Brian Bramson, George Sparlng, and I had devised for 'inverting' the standard twistor contour integral for massless free fields (Chapters 37 and 39), so that the twistor function is obtained directly from the field (see pp. [311]–[316]). There is also some discussion of massive fields, using a twistor and a dual twistor (compare Chapter 66); there are several ideas being put forward in relation to twistor diagrams (compare Chapter 52); and the asymptotic twistors of Chapter 52 are generalized to hypersurface twistors.

TWISTOR THEORY, ITS AIMS AND ACHIEVEMENTS

Roger Penrose

1. INTRODUCTION

In the physics of today there are two apparently independent places where a concept of continuum plays a fundamental role. First there is the real continuum of four dimensions, representing the arena of space-time within which the phenomena of the world are presumed to take place. This same continuum also appears in a different guise as the space of energy-momentum four-vectors. Secondly, there is the complex continuum of quantum mechanics, which gives rise to the concept of probability amplitude and the superposition law, leading to the picture of a complex Hilbert space for the description of phenomena. In the normal view, these two continua, one essentially real and the other essentially complex, are regarded as being quite independent of one another. And there seems no reason why quantum mechanics should not be applied within fictional space-times of dimension or signature different from that of our own - indeed, it often is! The ideas of quantum mechanics are simply superimposed on the classical picture of four-dimensional space-time and it seems that either continuum could have existed without the other. Apparently, from this point of view, there need be nothing special about the particular dimension and signature of that space-time in which we appear to live.

One of the basic ideas of twistor theory is to provide an alternative framework for physics in which, at the outset, these two continua are merged into one. The complex quantum mechanical continuum becomes inextricably woven into the geometry. The dimension and signature of space-time are now necessarily fixed; a twistor description would not work if these did not have the unique values that we observe for our physical universe. Thus, in effect, twistor theory reduces the number of distinct continuum concepts in physics from two to one. But a longer term aim of twistor theory is that ultimately the continuum concept may possibly be eliminated from the basis of physical theory altogether. Then the description of natural phenomena would be based on the principles of counting or other combinatorial processes.

There is a hint of an interrelation between the complex numbers of quantum mechanics and the structure of space even in the most elementary considerations of non-relativistic quantum mechanical spin(1)(2) The spin states of a spin $\frac{1}{2}$ particle are complex linear combinations of two states: 'spin up' and 'spin down'. But there is nothing special about the directions 'up' and 'down'. They might as well be 'left' and 'right', or 'forward' and 'backward'. So these other alternatives must be expressible as linear combinations of the first two. But to get the full array of directions in three-dimensional space it is necessary that complex combinations of the original two states be taken. Of course, things do get more complicated when systems of higher spin are considered. Nevertheless, the three-dimensionality of space and the need for complex numbers as probability amplitudes are clearly seen to be intimately related. This relationship finds its expression, in more precise mathematical terms, in the local isomorphism

$$SU(2) \to SO(3) \tag{1.1}$$

(which is a (2-1) map,cf.ref. (3)).

The representations of SO(3) are, indeed, best studied in their role as representations of SU(2). This fact, among other things, lies behind the combinatorial rules which form the basis of the theory of spin-networks(2). Since spin-networks form a model possessing features which, it is hoped, twistor theory will eventually be able to incorporate, I feel that it is worthwhile to digress briefly to discuss this theory. More complete accounts are to be found in refs. (2), (5) although the full arguments are as yet unpublished.

In this model, the physical world is pictured as a sort of network, an example of which is illstrated in Fig. 6.1.

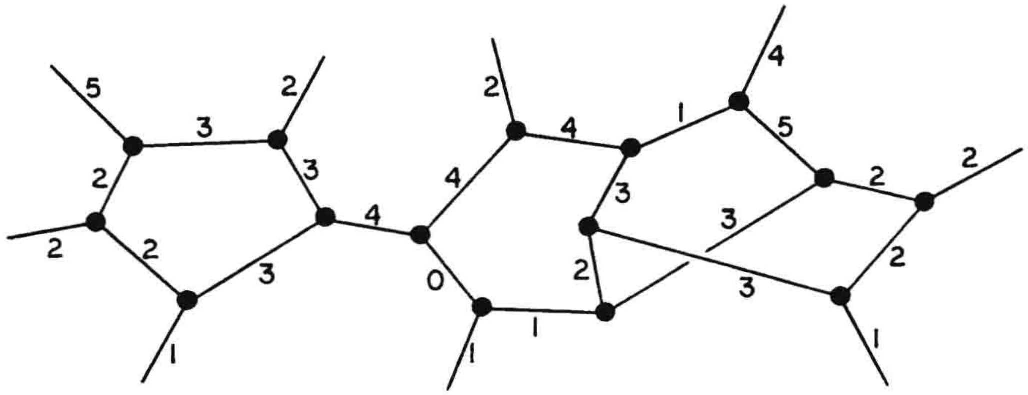


Fig. 6.1

Each line segment is to be thought of as representing an object having well-defined total angular momentum n/2 (taking $\hbar = 1$), where the (non-negative) integer n labels the line segment. The combinatorial network configuration and the values of these integers are all that is required to define the complete structure of the model. For purposes of visualization the line segments may be regarded as representing world-lines of slowly moving particles (although they need not be 'elementary particles' or anything so specific as that) and time may be viewed as proceeding from the bottom of the picture to the top - but the calculations are the same whichever direction time is taken in the diagram. Indeed, time and space do not enter into the structure at all at this stage. The idea is to construct a concept of physical space from the rules for the diagrams. These rules are purely combinatorial in nature and enable one to compute the probabilities for the various spin values which can occur when a new external line segment is added to a spin-network. (This must be joined to the network at one point where two previously external lines are brought together.) These probabilities turn out always to be rational numbers. They depend on the structure and spin values for the entire network and can be computed from certain integers which arise solely from combinatorial considerations of closed spin-network (i.e., with no free ends). Having these probability values, a concept of physical geometry can be built up by asserting that spin-network segments with large spin values define 'directions' (pictured as the spin axes) in the space, the 'angles' between these 'directions' being derived from the combinatorially computed probability values. Provided the spin-networks are such that these 'angles' are, in an appropriate sense, approximately 'rigidly defined', the geometry so obtained agrees (to a corresponding degree of approximation) with the geometry of directions in a Euclidean 3-space. On the other hand, if some of the 'angles' are not 'rigidly defined', we may get a number of co-existing approximate Euclidean 3-spaces with the property that 'rigidly defined' directions in one are, in a certain sense, linear superpositions of directions in the others. The picture which arises has strong analogies with that of a curved space, where different incompatible Euclidean geometries arise in an approximate way from each local region. (See ref (5) for a fuller discussion.)

As a description of the actual physical world, the spin-network model is deficient in two different (but related) ways: for the rules are such that the angular momentum is both non-relativistic and without any mixing of spin with orbital angular momentum. The latter deficiency implies that only directions arise in the resulting geometry and not spatial separations; and the former, that time is absent from the geometry. The obvious remedy for both deficiencies would be to attempt to replace SO(3) by the Poincaré group in the derivation of the combinatorial rules for the spin-networks. (To use merely the Lorentz groups would not incorporate spin-orbit mixing; for this we need the translations too). But this is no easy matter. There are several reasons, in fact, why it is more satisfactory to enlarge the group still further to the conformal group. This implies that particles are taken to have zero rest-mass in the first instance; rest-mass is then regarded as a secondary concept arising, as a feature connected with interactions (see §3; and for a related point of view, cf. ref. (6)). The local isomorphism$^{(3),(4)}$

$$SU(2,2) \rightarrow C_+^\uparrow(1,3) \tag{1.2}$$

(which is a (4 - 1) map) now takes the place of (1.1), $C_+^\uparrow(1,3)$ denoting the identity-connected component of the 15-parameter conformal group of (compactified) Minkowski space-time. The use of SU(2,2) rather than SU(2) entails some crucial differences. The most immediate of these comes from the fact that SU(2,2) is non-compact, so in contrast with SU(2) it possesses infinite-dimensional irreducible representations. Furthermore, it is certain infinite dimensional representations of SU(2,2) - namely those describing zero rest-mass particles - that we are particularly interested in. Thus it is by no means clear that the above programme, even if carried through successfully, would enable physical laws to be reduced to a set of combinatorial rules.

In spite of this, it is possible to view the twistor programme as an attempt to extend the spin-network model in the above way, the local isomorphism (1.2) forming the basis of the approach. Indeed, one of the main ideas of twistor theory is that space-time points are not initially present in the theory. Points are taken to be derived objects, the twistors themselves being more basic. Though a precise definition is left until §2, we may for the moment regard a twistor as representing a (classical) massless particle in free motion. (In fact twistors are not physical particles; this is only for visualization purposes.) To fix a space-time point R, we would consider all free motions of zero rest-mass particles through R. The space-time picture is thus that of the complete light cone of R.

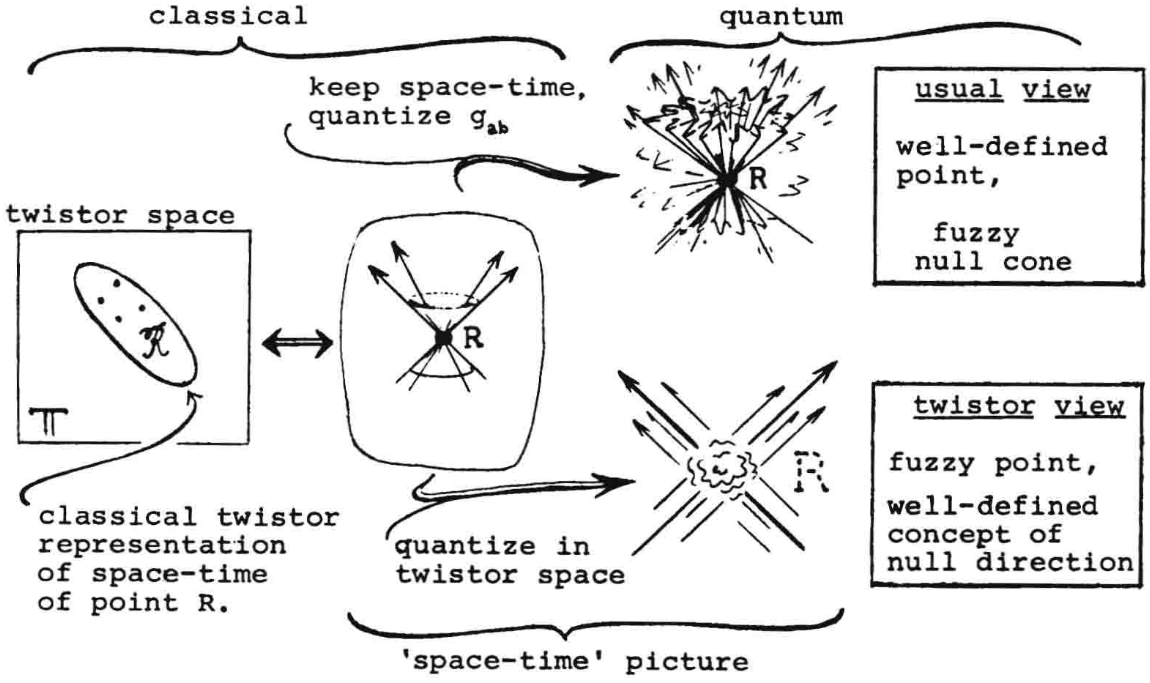

classical
quantum
keep space-time, quantize g_{ab}
R
usual view
well-defined point,
fuzzy null cone
twistor space
$\mathcal{R}$
$\mathbb{T}$
R
R
twistor view
fuzzy point,
well-defined concept of null direction
classical twistor representation of space-time of point R.
quantize in twistor space
'space-time' picture

Fig. 6.2

This light cone corresponds to a system $\mathcal{R}$ of twistors, so we have also a picture in twistor space $\mathbb{T}$, the original space-time point R being now represented by the subset $\mathcal{R}$ of $\mathbb{T}$.

While the geometrical description remains classical, these two pictures are really equivalent. But with the introduction of both curvature and quantum mechanics, the twistor description leads to a radically different view of the possible nature of a quantized space-time from the one normally presented. In the usual view, one imagines a space-time in which the points remain intact but where g_{ab} becomes quantized. Thus, the spacelike, null, or timelike nature of a vector at a point R becomes subject to quantum uncertainties. But according to the twistor view, things are the other way around. The space-time 'directions of motion' of the twistors must necessarily remain null; it is the property of space-time intersection between twistors which becomes subject to quantum uncertainties. The concept of a twistor remains intact (so the space $\mathbb{T}$ still exists), while that of a space-time point becomes fuzzy (see Fig. 6.2).

A viewpoint of this kind also fits in well with a belief (which is itself part of the twistor philosophy) that spinors are to be regarded as more fundamental than world-vectors - and that the signature (+ - - -) of our world arises in some way from the particularly simple nature of the associated spin-space.(1),(7). According to this belief, the relativistic generalization of (1.1), namely the local isomorphism(3),(4)

$$SL(2,C) \rightarrow O_{+}^{\uparrow}(1,3) \qquad (1.3)$$

(which is a (2 - 1) map), provides the 'reason' for the dimension and signature of space-time. The complex 2-dimensional spin-vectors (on which the SL(2,C) elements act when a Lorentz transformation is performed) are to be regarded as primary. All world-vectors and world-tensors are then to be taken as built up from spin-vectors by means of tensor products, sums and complex conjugations. But the space-time interpretation of a spin-vector is something locked on to the null cone (namely as a null flag, tangent to the cone). If it were uncertain whether a world-vector was null or not, then this uncertainty would have to extend to the very concept of a spin-vector. It is hard to see how the identity of an object as a spin-vector could be subject to quantum uncertainties if spin-vectors are to occupy the basic position in physics that this view requires. But the twistor view of the nature of space-time points provides a way out. The quantum uncertainties are transferred to the very concept of a space-time point.

One may still ask why so much faith has been placed in null objects (spin-vectors, zero rest-particles and fields, conformal invariance) in the twistor approach. After all, massive particles exist in nature (!) and physics seems not to be conformally invariant. It would be possible, indeed, to conceive a programme, initially similar to that of twistors, in which timelike geodesics (or geodesics in general) take over the role played by the null geodesics of twistor theory (compare ref. (8)). In a corresponding way, space-time points could still be considered as derived objects, the geodesic space being regarded as primary. The concept of a fuzzy point could again be entertained. Furthermore, the motivation from spin-network theory could be regarded as just as strong as for twistor theory. However, the value of any new approach must ultimately rest on its details. The strength of the twistor formalism resides in a number of remarkable mathematical facts which it is very hard to believe could be simply 'accidental'. These facts are all concerned with the complex structure of twistor space and the surprising and many-faceted roles played by holomorphic (i.e., complex-analytic) functions of twistors. The remarkable inter-relation between geometry and complex numbers that occurs in twistor theory is wholely dependent upon twistors being null objects; nothing like this would have arisen if the space of timelike geodesics had been used.

The complex numbers which define the structure of twistor space are intimately and inextricably tied in with the geometry of the resulting space-time. These same complex numbers emerge again in a different guise as quantum mechanical proability amplitudes. But even in the solution of purely classical relativistic field equations, holomorphic functions - where the real and imaginary parts of the functions are both significant - have played a surprisingly frequent role (c.f. ref (9) - (15). Twistor theory appears also to supply a partial explanation for this phenomenon. Such unifications and hints at unifications that the twistor approach provides are, to me, a stronger motivation than any of the more clear-cut achievements of the theory.

As an indication, in just one direction, of the present state of the actual achievements of twistor theory, it is worth mentioning that twistors do provide a very neat way of handling wave packets in scattering theory for zero rest-mass particles. Calculations are performed by integrating finite quantities over compact contours. The results of such integrations are always finite and it seems that the theory should be free of the usual divergences that arise in conventional quantum field theory. There is even the possibility that certain twistor integrals (those corresponding to closed twistor diagrams (cf. §4)) might be amenable to a combinatorial treatment. If so, the analogy with spin-network theory might turn out to be very close. However, it is as yet unclear precisely how all these finite integrals are to be interpreted or, indeed, how the twistor integrals are to be constructed, in general. Moreover, although the introduction of rest-mass into the theory can be achieved in certain ways, it seems one may then be led to integrals over non-compact contours, so the finiteness of the answers would no longer be guaranteed. It may well be the case that as soon as rest-mass is incorporated, general relativity must be incorporated simultaneously if there is to be a possibility of obtaining a finite theory. This would be in line with the hopes often expressed with regard to conventional theory, whereby quantized gravitation might eventually be able to supply the natural cut-off for all the divergent integrals which would otherwise arise(16)-(18). Unfortunately, at present these hopes remain little more than just that.

2. FLAT-SPACE TWISTORS

There are various different ways in which the twistor concept can be formally introduced. Perhaps the most directly physical of these is to use the momentum and angular momentum structure of a zero rest-mass particle. I shall be concerned only with the twistors for flat (or conformally flat) space-time here. Twistors for curved space-time will be discussed in §5.

Consider a classical system in Minkowski space $\mathbb{M}^I$ whose total momentum 4-vector* p^a and angular momentum bivector M^{ab} $(= - M^{ba})$, with

---

* Lower case Latin letters are used for 4-dimensional space-time indices; upper case Latin letters for 2-dimensional complex spinor indices, with a prime for the complex conjugate spin space. The indices used here are intended to be 'abstract' in the sense of ref. (7), but can in most places be interpreted without confusion in the conventional way as denoting sets of components. When numerical indices are used, these refer to a standard right-handed Minkowski coordinate system for which the squared line-element is $(dx^0)^2 - (dx^1)^2 - (dx^2)^2 - (dx^3)^2$ and a standard constant spin-frame according to which the vector V^a is represented by

$$\begin{pmatrix} V^{00'} & V^{01'} \\ V^{10'} & V^{11'} \end{pmatrix} = 2^{-\frac{1}{2}} \begin{pmatrix} V^0 + V^1 & V^2 + iV^3 \\ V^2 - iV^3 & V^0 - V^1 \end{pmatrix} .$$

Spinor indices are raised and lowered by

$$\alpha^A = \varepsilon^{AB}\alpha_B, \quad \alpha_B = \alpha^A\varepsilon_{AB}, \quad \beta^{A'} = \varepsilon^{A'B'}\beta_{B'}, \quad \beta_{B'} = \beta^{A'}\varepsilon_{A'B'}$$

$$\text{so} \quad \alpha^0 = \alpha_1, \quad \alpha^1 = -\alpha_0 , \quad \text{etc.}$$

Round brackets around indices denote symmetrization and square brackets, anti-symmetrization.

respect to some origin 0, are equivalent to those of a zero rest-mass particle of helicity s. We have

$$S_a = sp_a\ , \tag{2.1}$$

the vector

$$S_a = \tfrac{1}{2}\, e_{abcd}\, p^b\, M^{cd} \tag{2.2}$$

being the Pauli-Lubanski spin vector ($e_{abcd} = e_{[abcd]}$; $e_{0123} = 1$). Since p^a is null and future-pointing ($p_a p^a = 0$, $p^0 > 0$), the 2-spinor expression $p_{AA'}$ for p_a has the form

$$p_{AA'} = \bar{\pi}_A\, \pi_{A'} \tag{2.3}$$

for some (dual conjugate) spin-vector $\pi_{A'}$. The relation (2.1), with (2.2), then implies$^{(19)}$ that the angular momentum can be expressed as

$$M^{AA'BB'} = i\omega^{(A}\bar{\pi}^{B)}\,\varepsilon^{A'B'} - i\varepsilon^{AB}\,\bar{\omega}^{(A'}\pi^{B')} \tag{2.4}$$

for some spin-vector ω^A. Assuming $\pi_{A'} \neq 0$, the pair $(\omega^A, \pi_{A'})$ uniquely defines, via (2.3), (2.4), a momentum p_a and an angular momentum M^{ab} satisfying the above requirements for a massless particle. Conversely, given p_a and M^{ab} as above, the freedom in $(\omega^A, \pi_{A'})$ is just that of an overall phase change (θ real).

$$\omega^A \to e^{i\theta}\omega^A,\quad \pi_{A'} \to e^{i\theta}\,\pi_{A'}$$

Let us call ω^A and $\pi_{A'}$ the two spinor parts of the twistor* Z^α and write

$$Z^\alpha \leftrightarrow (\omega^A, \pi_{A'}) . \tag{2.5}$$

The complex conjugate of the twistor Z^α is a dual object $\bar{Z}_\alpha$ expressed as

$$\bar{Z}_\alpha \leftrightarrow (\bar{\pi}_A, \bar{\omega}^{A'}) \quad , \tag{2.6}$$

the scalar product of Z^α with $\bar{Z}_\alpha$ being

$$Z^\alpha \bar{Z}_\alpha = \omega^A \bar{\pi}_A + \pi_{A'} \bar{\omega}^{A'} . \tag{2.7}$$

Substituting (2.3) and (2.4) into (2.2) and (2.1) we get (after some calculation)

$$Z^\alpha \bar{Z}_\alpha = 2s , \tag{2.8}$$

so that the twistor 'norm' (2.7) measures the helicity of the zero rest-mass particle.

The complex conjugate rule for the twistor components

---

* The conventions used here follow those of ref. (19). They differ in certain essential respects from those of refs. (20) and (32).

$$(Z^0, Z^1, Z^2, Z^3) = (\omega^0, \omega^1, \pi_{0'}, \pi_{1'})$$

takes the form

$$\overline{Z^0} = \bar{Z}_2, \quad \overline{Z^1} = \bar{Z}_3, \quad \overline{Z^2} = \bar{Z}_0, \quad \overline{Z^3} = \bar{Z}_1 \qquad (2.9)$$

in a standard frame, where $\overline{Z^0}$ means the complex conjugate of the number Z^0, etc. The Hermitian form (2.8) can then be written

$$Z^0\overline{Z^2} + Z^2\overline{Z^0} + Z^1\overline{Z^3} + Z^3\overline{Z^1} = 2s, \qquad (2.10)$$

this having signature (+ + - -).

The explicit relations between the twistor components and those of the momentum and angular momentum are given by

$$\begin{aligned}
p^0 &= 2^{-\frac{1}{2}}(Z^3\bar{Z}_1 + Z^2\bar{Z}_0), \quad p^1 = 2^{-\frac{1}{2}}(Z^3\bar{Z}_1 - Z^2\bar{Z}_0) \\
p^2 &= 2^{-\frac{1}{2}}(-Z^2\bar{Z}_1 - Z^3\bar{Z}_0), \quad p^3 = i2^{-\frac{1}{2}}(Z^2\bar{Z}_1 - Z^3\bar{Z}_0) \\
M^{01} &= -M^{10} = \tfrac{i}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 - Z^2\bar{Z}_2 + Z^3\bar{Z}_3), \\
M^{02} &= -M^{20} = \tfrac{i}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 - Z^2\bar{Z}_3 - Z^3\bar{Z}_2), \\
M^{03} &= -M^{30} = \tfrac{1}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 - Z^2\bar{Z}_3 + Z^3\bar{Z}_2), \\
M^{12} &= -M^{21} = \tfrac{i}{2}(Z^0\bar{Z}_1 - Z^1\bar{Z}_0 + Z^2\bar{Z}_3 - Z^3\bar{Z}_2) \quad, \qquad (2.11) \\
M^{32} &= -M^{23} = \tfrac{1}{2}(Z^0\bar{Z}_0 - Z^1\bar{Z}_1 + Z^2\bar{Z}_2 - Z^3\bar{Z}_3), \\
M^{13} &= -M^{31} = \tfrac{1}{2}(Z^0\bar{Z}_1 + Z^1\bar{Z}_0 + Z^2\bar{Z}_3 + Z^3\bar{Z}_2), \\
&M^{00} = M^{11} = M^{22} = M^{33} = 0 \ .
\end{aligned}$$

The above representation of a twistor in terms of components is given for a particular choice of origin O. We shall investigate how the description of the twistor Z^α changes as the origin varies. The abstract twistor itself is to be thought of as a physical or geometrical object. It is not dependent upon this arbitrary choice of origin, but its representation in terms of components is. The (abstract) spinor parts of Z^α will also depend upon the choice of origin (although one of them only trivially). If X denotes an arbitrary 'origin-point' in $\mathbb{M}^I$, then ω^A and $\pi_{A'}$ will be functions of X. Indeed, we may think of ω^A and $\pi_{A'}$ as very special types of spinor field. To obtain their dependence on position we recall the transformation formula for angular momentum under change of origin. Let O be a fixed origin and X a variable point whose position vector with respect to O is x^a. Then, writing the symbol '0' beneath quantities defined at O, we have, for the momentum p_a and angular momentum M^{ab} about X:

$$p_a = \underset{0}{p}_a \quad , \quad M^{ab} = \underset{0}{M}^{ab} - 2x^{[a}\underset{0}{p}^{b]} .$$

Referring back to (2.3) and (2.4), and requiring that the phase of $\pi_{A'}$ be also constant (i.e., that the flag-planes of $\pi_{A'}$ be parallel) we obtain, for the spinor parts of Z^α at the general point X:

$$\omega^A = \underset{0}{\omega}^A - i\, x^{AA'} \underset{0}{\pi}_{A'} \quad , \quad \pi_{A'} = \underset{0}{\pi}_{A'} . \qquad (2.12)$$

Note that the spinor field ω^A depends linearly on position. Furthermore, it satisfies the 'field equation'

$$\nabla_{B'}^{(B}\omega^{A)} = 0 \tag{2.13}$$

where $\nabla_{AA'}$ denotes the ordinary gradient (or covariant derivative) operator $\partial/\partial x^{AA'}$. The equation (2.13) is called the twistor equation. It is not hard to show conversely$^{(19),(20)}$ that the general solution of (2.13) in fact has the form of the expression for ω^A in (2.12), where $\underset{0}{\omega}{}^A$ and $\underset{0}{\pi}{}_{A'}$ are arbitrary (constant) spinors. Thus we can use solutions of the twistor equation as an alternative way of defining twistors (in $\mathbb{M}^I$). We can extract the (constant) spinor field $\pi_{A'}$ directly from the field ω^A by

$$\nabla_{AA'}\omega^B = -\,i\,\varepsilon_A{}^B\,\pi_{A'}\ . \tag{2.14}$$

The pair of fields $(\omega^A,\pi_{A'})$, related by (2.14), can also be regarded as defining the twistor Z^α. The momentum and angular momentum, about each point space-time of the classical massless particle specified by Z^α are then described, according to (2.3) and (2.4), by the fields ω^A and $\pi_{A'}$.

One advantage of using the twistor equation (2.13) to define the concept of a twistor is that the conformal invariance is then explicit. Suppose the metric g_{ab} of $\mathbb{M}^I$ is rescaled according to

$$g_{ab} \to \hat{g}_{ab} = \Omega^2\, g_{ab} \quad , \tag{2.15}$$

where Ω is a smooth positive function on $\mathbb{M}^I$. The resulting metric

$\hat{g}_{ab}$ will in general be curved (although, of course, it must be conformally flat). We accompany (2.15) by

$$\hat{\varepsilon}_{AB} = \Omega\varepsilon_{AB},\ \hat{\varepsilon}_{A'B'} = \Omega\varepsilon_{A'B'},\ \hat{\varepsilon}^{AB} = \Omega^{-1}\varepsilon^{AB},\ \hat{\varepsilon}^{A'B'} = \Omega^{-1}\varepsilon^{A'B'} \quad . \tag{2.16}$$

(This is consistent with the 'abstract index' equation $g_{ab} = \varepsilon_{AB}\varepsilon_{A'B'}$. The components of $\hat{\varepsilon}_{AB}$ can be taken to be $0, \Omega$; $-\Omega, 0$ if desired, but the more convenient standard form $0, 1; -1, 0$ can also be retained if the basis spinors are also rescaled suitably). The covariant derivative operator $\hat{\nabla}_a$ for $\hat{g}_{ab}$ is obtained from the original ∇_a by

$$\hat{\nabla}_{AA'}\chi_{B\ldots G'}{}^{K\ldots Q'} = \nabla_{AA'}\chi_{B\ldots G'}{}^{K\ldots Q'} - T_{BA'}\chi_{A\ldots G'}{}^{K\ldots Q'} - \ldots - T_{AG'}\chi_{B\ldots A'}{}^{K\ldots Q'} +$$

$$+ \varepsilon_A{}^K T_{YA'}\chi_{B\ldots G'}{}^{Y\ldots Q'} + \ldots + \varepsilon_{A'}{}^{Q'} T_{AY'}\chi_{B\ldots G'}{}^{K\ldots Y'} \tag{2.17}$$

where

$$T_a = \Omega^{-1}\nabla_a\Omega = \nabla_a \log \Omega \tag{2.18}$$

Applying (2.17) to (2.13) we obtain

$$\hat{\nabla}_{B'}^{(B}\omega^{A)} = 0$$

showing that the twistor equation is conformally invariant and that the ω-field description of a twistor is independent of the conformal scaling.

CHAPTER 69

We can also apply (2.17) to (2.14) and derive the law

$$\hat{\omega}^A = \omega^A, \quad \hat{\pi}_{A'} = \pi_{A'} + i\, T_{AA'}\omega^A \tag{2.19}$$

for the behaviour of the $(\omega^A, \pi_{A'})$-description of a twistor Z^α under conformal rescaling. It should further be noted that

$$\hat{\omega}^A\hat{\bar{\pi}}_A + \hat{\pi}_{A'}\hat{\bar{\omega}}^{A'} = \omega^A\bar{\pi}_A + \pi_{A'}\bar{\omega}^{A'}$$

whence $2s = Z^\alpha\bar{Z}_\alpha$ is invariant under conformal rescaling, the measure of helicity, therefore, being a conformal invariant. (The independence of s under origin change is also manifest, from (2.12)).

If Ω has the particular form

$$\Omega = A\{(x^a - B^a)(x_a - B_a)\}^{-1} \quad , \tag{2.20}$$

with A and B^a constant, then g_{ab} will be again flat - although not defined on the light cone of the point B whose position vector is B^a. The action of the (restricted) Poincaré group with respect to these various conformally related flat metrics gives a set of transformations which generate the 15-parameter conformal group $C_+^\uparrow(1,3)$. The twistor equation is manifestly Poincaré invariant - and this is true for each of the above flat metrics. Hence, the twistor concept is invariant under this 15-parameter conformal group.

However there is an additional feature which must be taken into account when twistors and the conformal group are considered globally. We have seen that (2.20) is not defined on a certain light cone. In effect that light cone becomes the 'infinity' for the new flat-space metric. Indeed, the conformal group does not, strictly speaking, act on Minkowski space at all but on the *compactified* Minkowski space $\mathbb{M}$. This is obtained$^{(21),(20),(4)}$ from $\mathbb{M}^{I}$ by adjoining a (closed up) light cone, $\mathscr{I}$, at infinity and identifying across the cone*. The vertex of $\mathscr{I}$ is a new point, I, which from the point of view of the conformal structure of $\mathbb{M}$ is on an equal footing with every other point of $\mathbb{M}$. With a different choice of flat metric (say that given by (2.15) with Ω as in (2.20)) we would get a different point (in this case B) taking over the role of I as the vertex of the new light cone at infinity. The compact and conformally homogeneous space $\mathbb{M}$ has the topology $S^3 \times S^1$.

Owing to the conformal invariance of (2.13), we can locally extend any solution ω^A across $\mathscr{I}$ - simply by choosing another metric which is finite there. So we might expect that ω^A would be defined globablly over the whole of $\mathbb{M}$. However, it turns out that ω^A jumps by a factor i (or else - i, depending on the choice of spin-structure for $\mathbb{M}$) across $\mathscr{I}$. We may thus think of ω^A as a spinor field defined on the fourfold covering space(22) of $\mathbb{M}$ - or, equivalently, as a 'four-valued' field on $\mathbb{M}$, the four values at any one point being ω^A, $i\omega^A$, $-\omega^A$, $-i\omega^A$. This fourfold ambiguity can be attributed to the four-to-one nature of the local isomorphism (1.2). Indeed, the previous discussion can be seen to imply that under the action of $C^{\uparrow}_{+}(1,3)$ on $\mathbb{M}$, the twistors are transformed according to the group SU(2,2). The relevant invariant Hermitian form is simply (2.10).

---

* No *physical* significance is to be attributed to the fact that here future null infinity $\mathscr{I}^+$ and past null infinity $\mathscr{I}^-$ are being identified. Physical (massless) fields would generally be discontinuous across $\mathscr{I}$ whenever scattering takes place.

Let us now turn to the geometrical description of a twistor Z^α. There are two essentially different situations to consider, namely that of a null twistor ($Z^\alpha\bar{Z}_\alpha = 0$) and a non-null twistor ($Z^\alpha\bar{Z}_\alpha \neq 0$). In the first case (and assuming $Z^\alpha \neq 0$) there is a locus of points in $\mathbb{M}$ along which ω^A vanishes. This locus is a null straight line (or null geodesic, in case $\mathbb{M}$ is given a metric which is only conformally flat) which we label simply Z. We may think of Z physically as the world-line of the classical massless particle described by Z^α; it is the locus along which M^{ab} vanishes. The equation of Z is obtained simply by putting $\omega^A = 0$ in (2.12), i.e.,

$$\underset{O}{\omega}^{A} = ix^{AA'}\underset{O}{\pi}_{A'} \qquad (2.21)$$

x^a being the position vector, with respect to the origin O, of a point on Z (see Fig. 6.3). Having one such x^a, the others are obtained by replacing $x^{AA'}$ by $x^{AA'} + k\bar{\pi}^{A}\pi^{A'}$ for varying real k. This shows that Z is indeed a null straight line in the direction of the momentum p^a. The line Z defines the twistor Z^α completely up to a complex factor of proportionality. The modulus of this factor is determined by the momentum scaling (i.e., by the 'energy') of the massless particle; the argument by the phase of $\pi_{A'}$, which can be pictured geometrically$^{(7)}$ as a 'flag-plane' - a 2-plane element which is null and orthogonal to Z.

At each point X not on Z we have a non-zero spin-vector ω^A. This defines a null vector w^a by

$$w^{AA'} = \omega^A \bar{\omega}^{A'}. \qquad (2.22)$$

Now the null straight line through the point X in the direction of w^a intersects Z (possibly on $\mathcal{I}$). This is easily seen to be true at O, since a null x^a satisfying (2.21) must be proportional to w^a, and the general

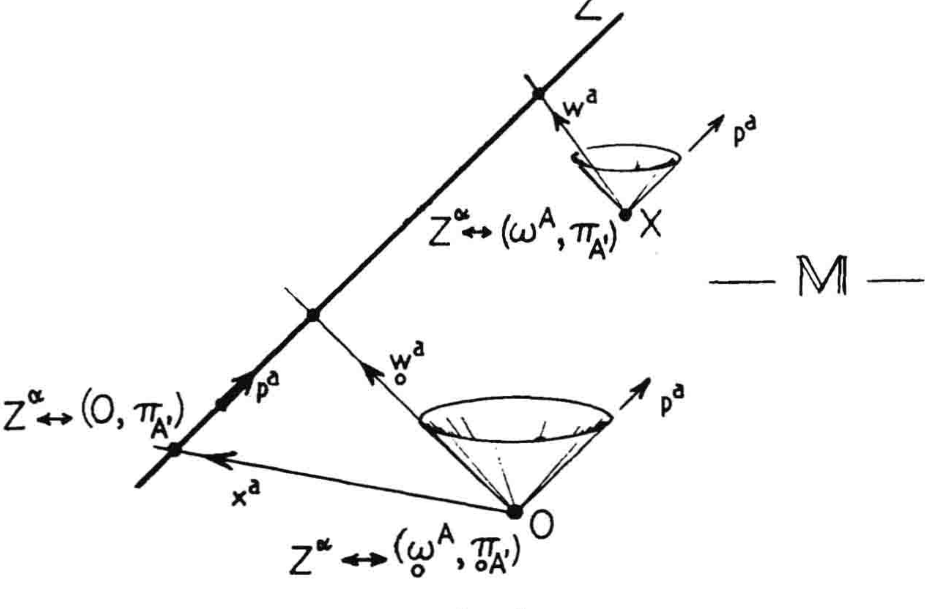

Z
w^a
p^a
$Z^\alpha \leftrightarrow (\omega^A, \pi_{A'})$
X
— M —
$\underset{o}{w}^a$
p^a
$Z^\alpha \leftrightarrow (0, \pi_{A'})$
p^a
x^a
O
$Z^\alpha \leftrightarrow (\underset{o}{\omega}^A, {}_o\pi_{A'})$

Fig. 6.3

result follows by translating the origin. Thus, the w-field defines the set of tangents to the congruence of null straight lines in $\mathbb{M}$ meeting Z. It should be noted that this still holds if $\pi_{A'} = 0$, even though the description of the twistor as defining the momentum and angular momentum of a massless particle now breaks down. In this case the ω^A field is constant (cf. (2.12)), so the w^a are all parallel. The line Z becomes one of the generators of *I*, so the massless particle is now at infinity - and has zero energy! The geometry still makes good sense, however, provided we use the compactified $\mathbb{M}$, rather than the original $\mathbb{M}^I$.

When $Z^\alpha \bar{Z}_\alpha \neq 0$ we can still use the w^a directions to obtain a geometrical realization of the twistor Z^α (up to proportionality) even though there are now no (real) points of $\mathbb{M}$ at which ω^A vanishes and so no (real) null line Z. The w-vectors are still tangents to a congruence of null straight lines in $\mathbb{M}$, but now they twist about one another* and never quite meet. The twist is in a right-handed sense if $Z^\alpha \bar{Z}_\alpha > 0$ and left-handed if $Z^\alpha \bar{Z}_\alpha < 0$. Such a congruence is termed a Robinson congruence and its detailed geometrical description(20) can be given as follows. Consider a space-like hyperplane section *S* of $\mathbb{M}^I$ given, say, by x^0 = const. At each point of the Euclidean 3-space *S*, the direction of the line of the congruence will be that of w^a, which we can visualize by projecting orthogonally into *S*. This defines a unit vector field v^a in *S*, given by

---

* This is the origin of the name "twistor".

$$(v^0, v^1, v^2, v^3) = (w^0)^{-1}\ (0, w^1, w^2, w^3) \qquad (2.23)$$

The geometry of v^a is fully determined by the location of a certain (oriented) circle γ in S (together with a choice of handedness for the congruence). There is a nested system of coaxal circular tori for which γ is the limiting member (Fig. 6.4).

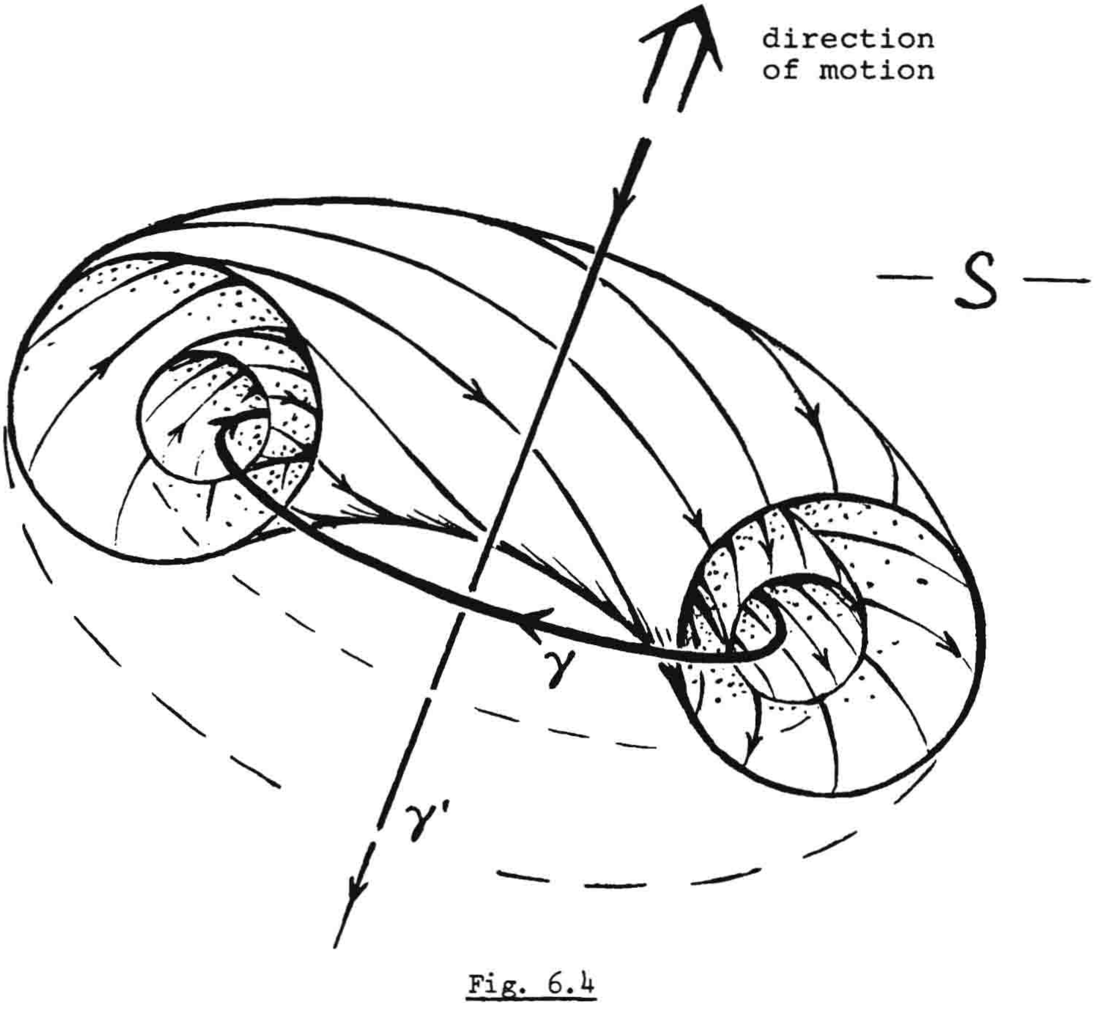


Fig. 6.4

On each torus the vector field v^a is tangent to a system of circles on the torus which link once through the torus and once about it. Thus S is covered by a system of (oriented) circles (including γ), every pair of which is linked - together with a straight line γ' through the centre of γ and perpendicular to its plane. There is a unique tangent direction, and thus a unique v^a, at each point of S. The lines of the Robinson congruence are consequently defined in terms of their intersections with S, their directions being determined by (2.23).

If we now let the time x^o progress, each line is described by a point moving rectilinearly with the speed of light in the direction of v^a. We find (rather remarkably) that the circle γ moves along with the speed of light along γ'. We may think of γ as a (relativistically non-invariant) characterization of the location and spin of the massless particle described by Z^α, its radius being a measure of the ratio of spin to energy. In the limit when $Z^\alpha\bar{Z}_\alpha = 0$, γ shrinks to a point, its space-time locus being simply Z.

It is worth remarking, at this stage, that the vector field w^a has another geometrical significance. It follows from (2.13) and (2.22) that w^a satisfies the conformal Killing equation

$$\nabla^{(a}w^{b)} = \frac{1}{4} g^{ab} \nabla_c w^c \quad . \tag{2.24}$$

Indeed, it can be shown that every null solution of (2.24) has the form (2.22) (or minus this form) for some ω^A satisfying (2.13). The angular momentum (2.4) can be obtained directly* by

---

* Hence, by taking linear combinations (cf. §3), we infer that the angular momentum structure of any system in special relativity has the form of the curl of a conformal Killing vector.

$$M^{ab} = \nabla^{[a}\, w^{b]} \qquad (2.25)$$

as follows from (2.14).

It may also be pointed out that the twistor equation (2.13) for the spinor field ω^A, when written in 4-spinor notation, takes the form

$$\{\gamma_a \nabla_b + \gamma_b \nabla_a - \tfrac{1}{2} g_{ab}\, \gamma^c\, \nabla_c\}\, \alpha \;= 0 \qquad (2.26)$$

where $\gamma_0, \gamma_1, \gamma_2, \gamma_3$ are the usual Dirac γ-matrices and α is a Dirac spinor defined so that the components of $(1 + i\gamma_5)\,\alpha$ are those of ω^A in a suitable frame and the components of $(1 - i\gamma_5)\,\alpha$ are those of another spinor $\tilde{\omega}^{A'}$ satisfying the conjugate of (2.13). If in addition we choose $(1 - i\gamma_5)\alpha = 0$, so that $\tilde{\omega}^{A'} = 0$, or else choose α to be a Majorana spinor, so that $\tilde{\omega}^{A'} = \bar{\omega}^{A'}$, then (2.26) is equivalent to (2.13). The original supersymmetry generators of Wess and Zumino(23) are objects of this kind and are therefore essentially twistors. (However Wess and Zumino required their solutions of (2.26) to be operators satisfying anticommutation relations.) The connection between solutions of (2.26) and conformal Killing vectors has also been noted by Wess and Zumino.

So far we have been concerned only with descriptions in real Minkowski space. However, much of twistor theory becomes greatly simplified once we allow ourselves the luxury of considering complexified space-time. I shall use the prefix $\mathbb{C}$ to denote the complexified version of a space. Thus, the points of $\mathbb{CM}^I$ can be labelled, with respect to the (real) origin O, by position vectors

$$z^a = x^a + i\, y^a \qquad (2.27)$$

where x^a and y^a are real world-vectors. It now turns out that for any twistor Z^α there is locus of points in $\mathbb{C}\mathbb{M}^I$ at which ω^A vanishes - except that if $\pi_{A'} = 0$ (and $\omega^A \neq 0$) the locus lies entirely on $\mathbb{C}\mathcal{I}$. This locus is obtained by replacing $x^{AA'}$ by $z^{AA'}$ in (2.21). Since I shall, from now on (until §5), be concerned with descriptions with respect to a fixed origin 0 only I shall, henceforth, drop the "0" from beneath the symbols and assume all relevant quantities to be defined at 0. Eqn. (2.21), for complex points, becomes

$$\omega^A = iz^{AA'}\pi_{A'} \quad . \tag{2.28}$$

Assuming $\pi_{A'} \neq 0$, the solutions form a 2-complex-dimensional locus (i.e. of 4 real dimensions) which is a totally null plane. For if z^a satisfies (2.28) then so also does

$$z^{AA'} + \lambda^A\pi^{A'} \tag{2.29}$$

where λ^A is arbitrary, the complex world vectors $\lambda^A\pi^{A'}$ being all null and orthogonal to one another. I shall denote this locus by $\mathbb{C}Z$ (for $Z^\alpha \leftrightarrow (\omega^A,\pi_{A'})$) even though this is not really the complexification of Z, whether $Z^\alpha\bar{Z}_\alpha = 0$ or not*. If $\pi_{A'} = 0$ (but $\omega^A \neq 0$) we still get a totally null complex plane, but now entirely on $\mathbb{C}\mathcal{I}$. If a point R of $\mathbb{C}\mathbb{M}$ lies on $\mathbb{C}Z$ we shall sometimes say that R *lies on the twistor* Z^α or that Z^α *passes through* R. Notice that the set of points of $\mathbb{C}Z$ defines Z^α uniquely up to porportionality ($Z^\alpha \neq 0$). Notice, further, that $\mathbb{C}Z$ contains a real point only if $Z^\alpha\bar{Z}_\alpha = 0$.

---

* The rationale for this is that the complex locus $\mathbb{C}Z$ is much more significant for the twistor Z^α than is the much 'smaller' real locus Z. Thus, Z is to be thought of as the set of real points of $\mathbb{C}Z$, rather than $\mathbb{C}Z$ as the complexification of Z.

If Z^α and Y^α are non-zero and non-proportional then $\mathbb{C}Z$ and $\mathbb{C}Y$ have a unique point in common. For if $Y^\alpha \leftrightarrow (\lambda^A, \mu_{A'})$ with $\mu_{A'}\pi^{A'} \neq 0$, this intersection point is obtained by solving $\lambda^A = iz^{AA'}\mu_{A'}$ and (2.28) simultaneously to obtain

$$z^{AA'} = i(\mu_{B'}\pi^{B'})^{-1}\ \{\omega^A\mu^{A'} - \lambda^A\pi^{A'}\} \tag{2.30}$$

On the other hand, if $\mu_{A'}\pi^{A'} = 0$, then either $\mathbb{C}Z$ is on $\mathbb{C}\mathcal{I}$ ($\mu_{A'} = 0$), or $\mathbb{C}Y$ is on $\mathbb{C}\mathcal{I}$ ($\mu_{A'} = 0$), or $\mathbb{C}Z$ and $\mathbb{C}Y$ are parallel ($\pi_{A'}, \mu_{A'}$ proportional); and in each case the intersection point is at infinity. The intersection point of $\mathbb{C}Z$ with $\mathbb{C}Y$ is *real* if and only if, together with $Z^\alpha\bar{Z}_\alpha = 0 = Y^\alpha\bar{Y}_\alpha$ we have

$$Z^\alpha\bar{Y}_\alpha = 0 \quad . \tag{2.31}$$

The set of non-zero X^α, for which $\mathbb{C}X$ passes through the intersection R of $\mathbb{C}Z$ with $\mathbb{C}Y$, have the form

$$X^\alpha = \rho Z^\alpha + \sigma Y^\alpha \tag{2.32}$$

for arbitrary ρ, σ, not both zero. Thus R itself is represented by

CHAPTER 69

the linear set (2.32). We can use the $\begin{bmatrix}2\\0\end{bmatrix}$ -twistor*

$$R^{\alpha\beta} = Z^{\alpha}Y^{\beta} - Y^{\alpha}Z^{\beta} \tag{2.33}$$

to represent the point R. The conditions for $R^{\alpha\beta}$ to be expressible in the form (2.33) are

$$R^{(\alpha\beta)} = 0, \qquad R^{[\alpha\beta}\,R^{\gamma]\delta} = 0 \ . \tag{2.34}$$

The second condition (2.34) (that $R^{\alpha\beta}$ be 'simple') can be expressed in various other equivalent forms, such as

$$R^{01}\,R^{23} + R^{02}\,R^{31} + R^{03}\,R^{12} = 0 \tag{2.35}$$

or

$$R^{\alpha\beta}\,R_{\alpha\beta} = 0 \qquad , \tag{2.36}$$

or

---

* A twistor of valence $\begin{bmatrix}p\\q\end{bmatrix}$, or $\begin{bmatrix}p\\q\end{bmatrix}$ -twistor, is denoted by a kernel symbol with p upper and q lower Greek indices. The usual algebraic tensor rules apply for the building up of higher valence twistors from the elements of the original twistor space (valence $\begin{bmatrix}1\\0\end{bmatrix}$), such as Z^{α}, and its dual (valence $\begin{bmatrix}0\\1\end{bmatrix}$), such as $\bar{Z}_{\alpha}$.

$$R^{\alpha\beta}\,\tilde{R}_{\alpha\gamma} = 0 \quad , \tag{2.37}$$

where the dual $\tilde{R}_{\alpha\beta}$ of $R^{\alpha\beta}$ is

$$\tilde{R}_{\alpha\beta} = \frac{1}{2}\,\varepsilon_{\alpha\beta\rho\sigma}\,R^{\rho\sigma} \tag{2.38}$$

whence

$$R^{\alpha\beta} = \frac{1}{2}\,\varepsilon^{\alpha\beta\rho\sigma}\,\tilde{R}_{\rho\sigma} \quad , \tag{2.39}$$

the alternating twistors $\varepsilon_{\alpha\beta\gamma\delta}$, $\varepsilon^{\alpha\beta\gamma\delta}$ being defined so that

$$\varepsilon_{\alpha\beta\gamma\delta} = \varepsilon_{[\alpha\beta\gamma\delta]} \quad , \quad \varepsilon^{\alpha\beta\gamma\delta} = \varepsilon^{[\alpha\beta\gamma\delta]} \quad , \quad \varepsilon_{0123} = \varepsilon^{0123} = 1 \tag{2.40}$$

(with components taken in the standard frame). The twistors $\varepsilon_{\alpha\beta\gamma\delta}$, $\varepsilon^{\alpha\beta\gamma\delta}$ are invariant under (active) SU(2,2) transformations (this being the essence of the "S" part of SU(2,2)); in particular, they are preserved by (2.12) and (2.19).

Defining the complex conjugate $\bar{T}^{\alpha\beta\gamma\ldots\eta}{}_{\rho\sigma\ldots\kappa}$ of a twistor $T_{\alpha\beta\gamma\ldots\eta}{}^{\rho\sigma\ldots\kappa}$ by the natural extension of (2.5), (2.6) and (2.9), so the standard components are related by

$$T_{320\ldots1}{}^{21\ldots0} = \bar{T}^{102\ldots3}{}_{03\ldots2} \quad , \tag{2.41}$$

CHAPTER 69

we have

$$\bar{\varepsilon}_{\alpha\beta\gamma\delta} = \varepsilon_{\alpha\beta\gamma\delta}\ , \quad \bar{\varepsilon}^{\alpha\beta\gamma\delta} = \varepsilon^{\alpha\beta\gamma\delta} \quad . \tag{2.42}$$

It then turns out that the complex conjugate $\bar{R}_{\alpha\beta}$ of $R^{\alpha\beta}$ gives a dual description of that point $\bar{R}$ of $\mathbb{CM}$ which is the complex conjugate of R (where if R has position vector z^a as in (2.27), $\bar{R}$ has position vector $\bar{z}^a = x^a - iy^a$). Furthermore, the dual of $\bar{R}_{\alpha\beta}$ is also the complex conjugate of $R_{\alpha\beta}$ (cf. (2.42)). Thus the notation $\bar{R}^{\alpha\beta}$ is unambiguous and denotes a twistor representing $\bar{R}$ in accordance with (2.33).

If R is a real point, $\bar{R}^{\alpha\beta}$ and $R^{\alpha\beta}$ are proportional. We can choose the phase of $R^{\alpha\beta}$ so that the proportionality constant is unity. Then, by (2.40) and (2.41), we obtain the reality restrictions on the components

$$\overline{R^{01}} = R^{01},\quad \overline{R^{23}} = R^{23},\quad \overline{R^{02}} = -R^{31},\quad \overline{R^{03}} = -R^{03},\quad \overline{R^{12}} = -R^{12} \tag{2.43}$$

from which it follows that the signature of (2.35) is (+ ---- +). The rays through the origin in the real 6-space that these components describe, subject to the restriction (2.35), give a model(21),(4) for the compactified Minkowski space $\mathbb{M}$. The SU(2,2) twistor transformations induce "rotations" in this 6-space, preserving the quadratic form (2.35), and give a realization of the local isomorphism(3),(4)

$$SU(2,2) \rightarrow O(2,4)$$

(which is (2-1) onto the connected component of O(2,4)), this being a kind of half way stage towards (1.2).

The point R need not be a finite point (i.e., in $\mathbb{M}^I$); indeed we can even take R to be the vertex I of $\mathcal{I}$. This gives dually related twistors $I^{\alpha\beta}$, $I_{\alpha\beta}$, which may be chosen explicitly as

$$I^{\alpha\beta} \longleftrightarrow \begin{bmatrix} \varepsilon^{AB} & 0 \\ 0 & 0 \end{bmatrix} , \quad I_{\alpha\beta} \longleftrightarrow \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon^{A'B'} \end{bmatrix} \qquad (2.44)$$

and which are then called the infinity twistors. The SU(2,2) transformations which preserve the point I (i.e., which preserve $I^{\alpha\beta}$ and $I_{\alpha\beta}$ up to proportionality) give the (restricted) Poincaré group together with dilations on $\mathbb{M}^I$. The dilations are eliminated by the requirement that $I^{\alpha\beta}$ is itself preserved; equivalently, that $I_{\alpha\beta}$ is preserved. Thus, a way of expressing the breaking of conformal invariance (while retaining Poincaré invariance), in twistor theory, is to incorporate the quantities $I^{\alpha\beta}$ and $I_{\alpha\beta}$ explicitly into the algebra. Note that, from (2.37), (2.38), (2.39), we have

$$I_{\alpha\beta} = - I_{\beta\alpha} = \bar{I}_{\alpha\beta} = \frac{1}{2} \varepsilon_{\alpha\beta\rho\sigma} I^{\rho\sigma} ,$$

$$I^{\alpha\beta} = - I^{\beta\alpha} = \bar{I}^{\alpha\beta} = \frac{1}{2} \varepsilon^{\alpha\beta\rho\sigma} I_{\rho\sigma} ,$$

$$I^{\alpha\beta} I_{\alpha\gamma} = 0 \quad . \qquad (2.45)$$

CHAPTER 69

We may use $I^{\alpha\beta}$, $I_{\alpha\beta}$ to give a normalization of $R^{\alpha\beta}$, $R_{\alpha\beta}$ in the cases where R is a finite point (real or complex) by taking

$$R^{\alpha\beta} I_{\alpha\beta} = 2; \quad \underline{\text{equivalently}} \quad R_{\alpha\beta} I^{\alpha\beta} = 2. \tag{2.46}$$

Then if r^a is the position vector of R with respect to O, we have by (2.30) and (2.33)

$$R^{\alpha\beta} \leftrightarrow \begin{bmatrix} -\frac{1}{2} r_c r^c \varepsilon^{AB} & i r^A{}_{B'} \\ -i r_{A'}{}^B & \varepsilon_{A'B'} \end{bmatrix}, \quad R_{\alpha\beta} \leftrightarrow \begin{bmatrix} \varepsilon_{AB} & i r_A{}^{B'} \\ -i r^{A'}{}_B & -\frac{1}{2} r_c r^c \varepsilon^{A'B'} \end{bmatrix}. \tag{2.47}$$

With this normalization, and if q^a is related to $Q^{\alpha\beta}$ in the same way that r^a is related to $R^{\alpha\beta}$, we have

$$Q^{\alpha\beta} R_{\alpha\beta} = -(q^a - r^a)(q_a - r_a). \tag{2.48}$$

Note that this implies that $Q^{\alpha\beta} R_{\alpha\beta} = 0$ if and only if the points Q and R have a null separation. This last fact does not require the normalization (2.46); indeed a point R of $\mathbb{C}\mathscr{I}$ is characterized by the fact that $I^{\alpha\beta} R_{\alpha\beta} = 0$, since R is then null-separated from I.

For the purposes of geometrical visualization it is often more convenient to work with the 3-complex-dimensional projective twistor space $\mathbb{P}\mathbb{T}$, rather than the 4-complex-dimensional twistor space $\mathbb{T}$.

Each point $\underset{\sim}{Z} \in \mathbb{PT}$ is the equivalence class of twistors obtained by multiplying a given non-zero twistor Z^α by all non-zero complex numbers (so Z^α and λZ^α give the same point $\underset{\sim}{Z}$ of $\mathbb{PT}$, $\lambda \neq 0$). From the above discussion we see that there is a (1-1) geometrical correspondence* between points of $\mathbb{CM}$ and the straight lines (complex projective lines) in $\mathbb{PT}$. The unique point R of intersection of $\mathbb{C}$ Z with $\mathbb{C}$ Y is represented as the unique line $\underset{\sim}{R}$ in $\mathbb{PT}$ joining the points $\underset{\sim}{Z}$ and $\underset{\sim}{Y}$ (see fig. 6.5). There is also a special line $\underset{\sim}{I}$ in $\mathbb{PT}$ representing the point I of $\mathbb{CM}$. The points of $\mathbb{C}\mathcal{I}$ in $\mathbb{CM}$ are represented, in $\mathbb{PT}$, by the lines meeting the line $\underset{\sim}{I}$.

To describe the real structure $\mathbb{M}$, we must invoke the (pseudo-) Hermitian correspondence $Z^\alpha \mapsto \bar{Z}_\alpha$. This induces a (pseudo-) Hermitian correlation on $\mathbb{PT}$, whereby points (described

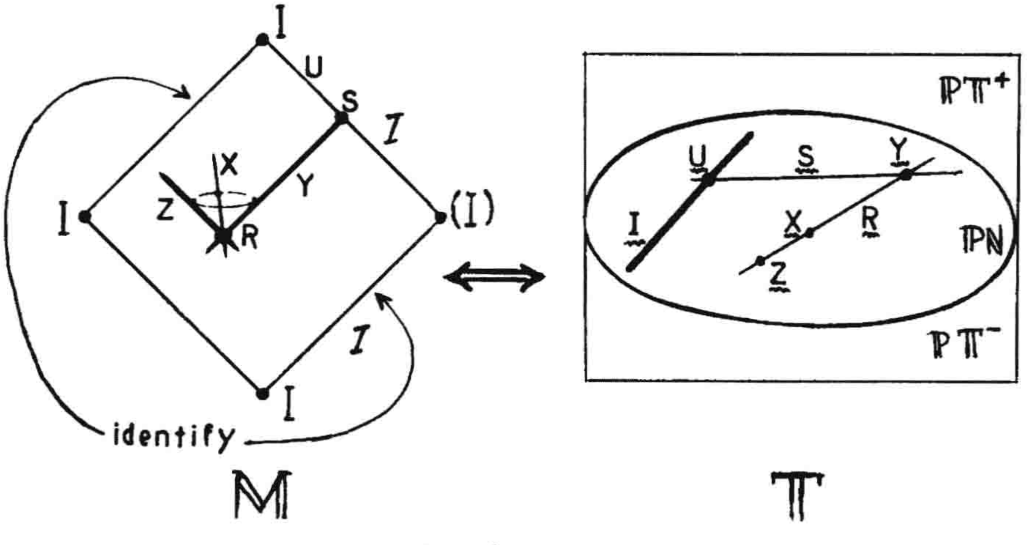

I
U
S
$\mathcal{I}$
X
Y
Z
R
I
(I)
$\mathcal{I}$
I
identify
$\mathbb{M}$
$\mathbb{PT}^+$
$\underset{\sim}{U}$
$\underset{\sim}{S}$
$\underset{\sim}{Y}$
$\underset{\sim}{I}$
$\underset{\sim}{X}$
$\underset{\sim}{R}$
$\mathbb{PN}$
$\underset{\sim}{Z}$
$\mathbb{PT}^-$
$\mathbb{T}$

Fig. 6.5

---

*This is a correspondence familiar to geometers: the space $\mathbb{CM}$ would be referred to as the Klein representation(24) of lines in $\mathbb{PT}$.

by twistors Z^α) and planes (described by twistors W_α) are interchanged. The image of the point $\underset{\sim}{Z}$ is plane $\bar{\underset{\sim}{Z}}$; and that of the plane $\underset{\sim}{W}$ is the point $\bar{\underset{\sim}{W}}$. (The plane $\underset{\sim}{W} \subset \mathbb{PT}$ is the set of points $\underset{\sim}{X}$ such that $W_\alpha X^\alpha = 0$, etc.) The point $\underset{\sim}{Z}$ lies on $\underset{\sim}{W}$ if and only if $\bar{\underset{\sim}{W}}$ lies on $\bar{\underset{\sim}{Z}}$. If $\underset{\sim}{R}$ is the line joining points $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$, then $\bar{\underset{\sim}{R}}$ is the intersection of the planes $\bar{\underset{\sim}{Y}}$ and $\bar{\underset{\sim}{Z}}$. The point of $\mathbb{CM}$ corresponding to $\bar{R}$ is then the complex conjugate of the point of $\mathbb{CM}$ corresponding to R.

A non-zero twistor Z^α is *null* ($Z^\alpha\bar{Z}_\alpha = 0$) if and only if the point $\underset{\sim}{Z} \in \mathbb{PT}$ lies on $\bar{\underset{\sim}{Z}}$. The set of such points is denoted by $\mathbb{PN}$, where $\mathbb{N}$ is the set of null twistors in $\mathbb{T}$. The set of points $\underset{\sim}{Z}$ in $\mathbb{PT}$ for which $Z^\alpha\bar{Z}_\alpha > 0$ is denoted $\mathbb{PT}^+$; the set for which $Z^\alpha\bar{Z}_\alpha < 0$ is denoted $\mathbb{PT}^-$. Thus the region $\mathbb{PN}$ (topology $S^3 \times S^2$) divides $\mathbb{T}$ (topology $\mathbb{C}P^3$) into two regions $\mathbb{PT}^+$ and $\mathbb{PT}^-$ (each of topology $S^2 \times E^3$). The points of $\mathbb{PN}$ represent the null lines of $\mathbb{M}$.

The plane $\underset{\sim}{W}$ (described by W_α) touches $\mathbb{PN}$ if and only if $W_\alpha\bar{W}^\alpha = 0$, in which case the contact point is $\bar{\underset{\sim}{W}}$. In any case, the intersection of $\underset{\sim}{W}$ with $\mathbb{PN}$ gives a system of points $\underset{\sim}{X}$ (described by $W_\alpha X^\alpha = 0 = \bar{X}_\alpha X^\alpha = 0$) defining a congruence of null lines X in $\mathbb{M}$. This is the Robinson congruence defined by the ω^A-field of $Z^\alpha = \bar{W}^\alpha$, which in the special case $\bar{W}^\alpha W_\alpha = 0$ degenerates to the system of null lines in $\mathbb{M}$ meeting the null line $\bar{W}$. A line $\underset{\sim}{R}$ in $\mathbb{PT}$ represents a real point of $\mathbb{CM}$ (i.e., a point of $\mathbb{M}$) if $\underset{\sim}{R}$ lies entirely in $\mathbb{PN}$ This is the case for which the imaginary part y^a of the position vector of the point vanishes. On the other hand, y^a ($\neq 0$) is space-like, null,or timelike according as $\underset{\sim}{R}$ cuts across $\mathbb{PN}$, touches $\mathbb{PN}$, or fails to intersect $\mathbb{PN}$. In the latter two cases, y^a is future-pointing or past-pointing according as $\underset{\sim}{R} \subset \mathbb{PN} \cup \mathbb{PT}^-$ or $\underset{\sim}{R} \subset \mathbb{PN} \cup \mathbb{PT}^+$.

The set of points $\mathbb{C}Z$ in $\mathbb{CM}$, for $\underset{\sim}{Z} \in \mathbb{PT}$ corresponds to the star of lines in $\mathbb{PT}$ through $\underset{\sim}{Z}$. The complex conjugate $\mathbb{C}\bar{Z}$ of $\mathbb{C}Z$, on the other hand, corresponds to quite a different kind of system of lines in $\mathbb{PT}$, namely the lines lying on a plane $\underset{\sim}{\bar{Z}}$. Thus, we have two kinds of totally null planes in $\mathbb{CM}$, those like $\mathbb{C}Z$ arising from $\begin{bmatrix}1\\0\end{bmatrix}$-twistors and those like $\mathbb{C}\bar{Z}$ arising from $\begin{bmatrix}0\\1\end{bmatrix}$-twistors. In the first case they are described as in (2.29) where $\pi^{A'}$ is held fixed and λ^A is varied. In the second case, λ^A is held fixed and $\pi^{A'}$ varied. Two distinct totally null planes of the same kind always have a unique point of $\mathbb{CM}$ in common. (The join of two points in $\mathbb{PT}$ is a unique line; the intersection of two planes in $\mathbb{PT}$ is a unique line.) However, two totally null planes of opposite kinds do not usually intersect at all; but when they do, they have an entire complex null line in common. (A general plane in $\mathbb{PT}$ contains no line through a general point in $\mathbb{PT}$; but if the point lies on the plane, there is a one-complex-parameter family of such lines.) Thus we can say that twistors Z^α and X^α always intersect. So also do twistors W_α and U_α. But Z^α and W_α will only intersect if $Z^\alpha W_\alpha = 0$, in which case they have an entire complex null line of points of $\mathbb{CM}$ in common.

3. TWISTOR DESCRIPTION OF FIELDS

In §2 we saw how the geometry of Minkowski space could be reformulated in terms of the geometry of twistor space. Although the two descriptions of the same object, say of a space-time point or a null line, might seem very different, the existence of a clear-cut translation procedure shows that the two pictures are, in a well-defined sense, equivalent. But this is not to say that the translation from one formalism to the other need be at all easy. Indeed, an advantage of having an alternative formalism can be that different operations may now be the ones that suggest themselves as being easy to perform. It is often the case that a direct translation into twistor terms of some concept which is familiar in Minkowski space may at first appear very elaborate and unnatural. But such elaborate expressions frequently admit of a surprising reformulation simpler even than those Minkowskian ones from which they arose. In these reformulations holomorphic functions of twistors play a key role.

Perhaps the most striking example of this phenomenon occurs with the zero rest-mass free-field equations. These may be expressed(25)-(28) as

$$\nabla^{AA'}\phi_{AB\ldots L} = 0 \tag{3.1}$$

where the symmetric spinor $\phi_{AB\ldots L}$ $(= \phi_{(AB\ldots L)})$ has n indices, the spin of the field being $\frac{1}{2}n$ with units chosen so that $\hbar = 1$. Strictly speaking, for positive frequency solutions of (3.1) the helicity is

$$s = -\tfrac{1}{2}n \quad . \tag{3.2}$$

On the other hand, the conjugate equation

$$\nabla^{AA'}\theta_{A'B'\ldots K'} = 0 \qquad (3.3)$$

with $\theta_{A'\ldots K'}$ $(= \theta_{(A'\ldots K')})$ having n' primed indices, has positive frequency solution with helicity

$$s = \frac{1}{2}n' \quad . \qquad (3.4)$$

If the spin is zero we must use a second-order equation

$$\nabla^a\nabla_a \phi = 0 \quad . \qquad (3.5)$$

The equations (3.1) and (3.3) are invariant under conformal rescalings (2.15), (2.16) where we choose

$$\hat{\phi}_{A\ldots L} = \Omega^{-1}\phi_{A\ldots L}\ , \quad \hat{\theta}_{A'\ldots K'} = \Omega^{-1}\theta_{A'\ldots K'} \qquad (3.6)$$

as easily follows from (2.17). To achieve conformal invariance for (3.5) we must, for the case of a curved space-time, replace (3.5) by $(\nabla^a\nabla_a + R/_6)\phi = 0$ where R is the Ricci scalar (and take $\hat{\phi} = \Omega^{-1}\phi$).

Equation (3.1) has various interesting special cases. If n = 1 we obtain the Weyl neutrino equation $\nabla^{AA'}\phi_A = 0$. If n = 2, (3.1) becomes the spinor version of Maxwell's source-free equations on the tensor F_{ab}, where[29]

$$F_{AA'BB'} = \phi_{AB}\varepsilon_{A'B'} + \varepsilon_{AB}\bar{\phi}_{A'B'} \quad . \qquad (3.7)$$

If n = 4 it becomes(28), with

$$K_{AA'BB'CC'DD'} = \phi_{ABCD}\varepsilon_{A'B'}\varepsilon_{C'D'} + \varepsilon_{AB}\varepsilon_{CD}\bar{\phi}_{A'B'C'D'} \quad , \qquad (3.8)$$

the "gauge invariant" version of the weak-field limit of Einstein's vacuum equations,(26),(30) the linearized curvature tensor K_{abcd} being defined as a field in Minkowski space. This satisfies

$$K_{abcd} = K_{[ab][cd]} = K_{cdab}, \quad K_{[abc]d} = 0, \quad K^{a}{}_{bad} = 0$$

by virtue of the form of the expression (3.8), the equation (3.1) now becoming the linearized Bianchi identity

$$\nabla_{[a} K_{bc]de} = 0$$

and can be taken as the integrability relation for a small change in the metric $g_{ab} \mapsto g_{ab} + \varepsilon h_{ab}$ (ε infinitesimal; $h_{ab} = h_{ba}$) to satisfy Einstein's vacuum equations, where

$$\phi_{ABCD} = \frac{1}{2} \nabla^{C'}_{(A} \nabla^{D'}_{B} h_{CD)C'D'}$$

The problem now is to find a natural way of expressing (3.1) (and also (3.3) and (3.5)) in twistor terms. The direct translation presents some unpleasant features. Fields are, after all, functions of points, but space-time points are not so easily represented in the twistor formalism. We might envisage using the description in terms of $R^{\alpha\beta}$ (cf. (2.47)), but the identical relations (2.35) pose something of a problem, especially when derivatives with respect to $R^{\alpha\beta}$ are to be contemplated. An alternative would be to use a pair of twistors to define the point, such the Z^{α} and Y^{α} of (2.33), but it seems complicated to consider "two-point" functions of twistors for even the simplest types of field - and furthermore these functions would have to satisfy certain differential equations, expressing the fact that they depend only on the combination (2.33) and not on the individual twistors Z^{α} and Y^{α}. These considerations apply already to scalar fields in $\mathbb{M}^{I}$. But, in addition, there is the spinor character of $\phi_{A\ldots L}$ (i.e., its index structure) to worry about, not to mention the coding of the specific field equations (3.1).

However, twistors are natural objects for the description of conformally invariant concepts in $\mathbb{M}$. And we have seen that the equations (3.1) are conformally invariant. This leads one to hope that something better can be done than merely directly translating (3.1) into twistor terms. There is a further point to note in this connection. Although the field $\phi_{A\ldots L}$ is defined throughout the four dimensions of space-time, the field equations (3.1) tell us that the entire field can be developed from components defined on a three-dimensional initial hypersurface. Thus, in effect, $\phi_{A\ldots L}$ is a function of only three variables. But the projective twistor space is three-dimensional (albeit of three complex dimensions). So one could hope that a description of the field $\phi_{A...L}$ might conceivably be obtainable in terms of homogeneous twistor functions of a single twistor variable (which should be holomorphic i.e., complex-analytic; otherwise $\mathbb{PT}$ would have to count as a six-dimensional space).

It turns out that such hopes are indeed justified(31),(32). We can describe solutions of (3.1) in terms of holomorphic functions of a twistor W_α, which are homogeneous of degree $-n-2$:

$$f(\lambda W_\alpha) = \lambda^{-n-2} f(W_\alpha) \quad . \tag{3.9}$$

To obtain $\phi_{A...L}$ from $f(W_\alpha)$ at a (real or complex) space-time point X, with position vector x^a, we perform a contour integral in the (dual) spin-space at X - which may be interpreted as the space of $\begin{bmatrix}0\\1\end{bmatrix}$ - twistors $W_\alpha \longleftrightarrow (W_A, -ix^{AA'}W_A)$ through X. Explicitly:

$$\phi(x^P)_{AB...L} = \frac{1}{2\pi i} \oint W_A W_B \ldots W_L \; f(W_P, -ix^{PP'}W_P) \; \Delta W \tag{3.10}$$

where

$$\Delta W = \varepsilon^{AB} \, W_A \, dW_B \quad . \tag{3.11}$$

The homogeneity (3.9) ensures that the exterior derivative (for constant x^a) of the integrand in (3.10) vanishes. It follows that (3.10) is a genuine contour integral, whose value is unaltered if the location of the (closed) contour is changed continuously, the contour necessarily avoiding the singularities of f. It is clear that f must indeed possess singularities (if $f \neq 0$) since it is holomorphic of negative homogeneity degree. (Thus the domain of f is not the whole of dual twistor space but is some appropriate subset of it.) I shall suppose that the singularities of f are suitably arranged so that non-trivial contours indeed exist and can be chosen to move suitably continuously as x^a varies. (In fact many - and in a sense sufficient - such f's can be constructed.) The fact that the $\phi_{A\ldots L}$ defined by (3.10) satisfies (3.1) is now a simple matter to verify directly.

Solutions of (3.3) are similarly constructed:

$$\theta(x^P)_{A'B'\ldots K'} = \frac{1}{2\pi i} \oint Z_{A'}Z_{B'}\ldots Z_{K'}\ g(ix^{PP'}\ Z_{P'},\ Z_{P'})\ \Delta Z \qquad (3.12)$$

where

$$g(\lambda Z^\alpha) = \lambda^{-n'-2}\ g(Z^\alpha); \qquad (3.13)$$

(with $Z = \varepsilon^{A'B'}\ Z_{A'}dZ_{B'}$), and solutions of (3.5) can be expressed in either of the two forms (3.10), (3.12), with $n = n' = 0$:

$$\phi(x^P) = \frac{1}{2\pi i} \oint f(W_P, -ix^{PP'}W_P)\ \Delta W\ \underline{\text{or}}\ \frac{1}{2\pi i} \oint g(ix^{PP'}\ Z_{P'},\ Z_{P'})\ \ \Delta Z. \ (3.14)$$

Again, the direct verification is a simple matter.

We ask how general are the fields obtained by this construction and, related to this, is there an inverse expression to (3.10), (3.12), (3.14) which obtains the twistor functions f, g from the space-time fields? A few preliminary remarks should be made. In the first place, the fields constructed by (3.10),(3.12), (3.14) must be analytic in x^a. This may seem to be a drawback from the physical point of view, but I do feel that this is really so. Non-analytic fields can be approximated arbitrarily closely by analytic ones and, furthermore, the positive and negative frequency parts of a field (if we are considering free waves) can be represented as boundary values of fields holomorphic in certain complex domains.(33) Even distribution-valued fields can be described in terms of holomorphic fields from the standpoint of hyperfunction theory(34). Indeed, it may well be that significant developments in twistor theory will come about by the systematic employment of hyperfunctions in place of the more usual δ-functions which appear in conventional field theory*.

As regards finding an inverse to (3.10), we should observe that there will be many f's giving rise to the same $\phi_{AB\ldots L}$. For if we replace f according to

$$f \mapsto f + h \qquad (3.15)$$

where h is non-singular inside (or on one side of) the contours

---

[*] I am grateful to P.D.F. Ion for acquainting me with the basic ideas of hyperfunction theory.

then the resulting field $\phi_{AB\ldots L}$ will be unchanged. Thus (3.15) gives us a sort of 'gauge freedom' for the twistor function, but it is not like the usual type of gauge transformation, since the 'gauge group' depends on the location of the contours. We cannot expect a unique inverse to (3.15), but one where there is a considerable arbitrariness involved. In fact, a recent development* in the theory, carried out in conjunction with B.D. Bramson and G.A.J. Sparling, appears to have provided a neat solution to the inverse problem, in which this arbitrariness emerges in a surprising way.

To express the solution, let me first reformulate (3.10) slightly. Choose a spinor basis (o^A, ι^A) at X normalized so that

$$o_A \iota^A = 1 \tag{3.16}$$

and let U_α and V_α be twistors through X corresponding to this basis:

$$U_\alpha \leftrightarrow (o_A, -ix^{AA'} o_A), \quad V_\alpha \leftrightarrow (\iota_A, -ix^{AA'} \iota_A). \tag{3.17}$$

Then (3.16) asserts (cf. (2.44)):

$$U_\alpha V_\beta I^{\alpha\beta} = 1 \quad . \tag{3.18}$$

The point X itself is also defined by U_α and V_α (via $X_{\alpha\beta} = 2U_{[\alpha}V_{\beta]}$, cf. (2.33)), so the components

---

[*] This development occurred after the lecture, on which these notes are based, was given.

$$\phi_r = \phi_{A\ldots ED\ldots L}\, o^A \ldots o^E \iota^D \ldots \iota^L \qquad (r = 0,\ldots,n) \tag{3.19}$$

may be regarded as just functions of U_α and V_β. Indeed, we have, taking components of (3.10) with respect to o^A and ι^A,

$$\phi_r(U_\alpha, V_\alpha) = \frac{1}{2\pi i} \oint \lambda^r\, f(\lambda U_\alpha - V_\alpha)\, d\lambda \;\; (r = 0,\ldots,n) \tag{3.20}$$

where $W_\alpha = \lambda U_\alpha - V_\alpha$ so that $W_A o^A = 1$, $W_A \iota^A = \lambda$ and $\Delta W = d\lambda$ (at fixed x^a). Eqn. (3.20) is a version of (3.10) which is expressed entirely in twistor terms. We may regard (3.20) as defining $\phi_r(U_\alpha, V_\alpha)$ even when the normalization (3.18) is dropped. Since this normalization may be viewed as just making a particular choice of conformal factor at X (cf. (2.16)), the effect is simply to incorporate the scaling behaviour (3.6) into (3.10).

From (3.20) we readily obtain

$$\frac{\partial \phi_r}{\partial U_\alpha} + \frac{\partial \phi_{r+1}}{\partial V_\alpha} = 0 \quad , \tag{3.21}$$

$$\frac{\partial^2 \phi_r}{\partial U_{[\alpha} \partial V_{\beta]}} = 0 \tag{3.22}$$

and

$$U_\alpha \frac{\partial \phi_r}{\partial U_\alpha} = -(r+1)\,\phi_r \;, \qquad V_\alpha \frac{\partial \phi_r}{\partial V_\alpha} = (r - n - 1)\,\phi_r \;. \tag{3.23}$$

We may think of (3.21) and (3.22) as twistor analogues of the zero rest-mass field equations (3.1) and of the wave equation, respectively, where the relations (3.23), which state that ϕ_r has homogeneity degrees $(-r-1,\ r-n-1)$, must also hold$^{(32)}$. In fact, (3.21) and (3.23) are together equivalent to (3.1), with (3.6).

The quantity

$$\phi(U_\alpha, V_\alpha) = \phi_o(U_\alpha, V_\alpha) = \frac{1}{2\pi i} \oint f(\lambda U_\alpha - V_\alpha)\, d\lambda \tag{3.24}$$

has an especial interest. It satisfies

$$\phi(\nu U_\alpha, V_\alpha + \mu\, U_\alpha) = \nu^{-1}\, \phi(U_\alpha, V_\alpha) \tag{3.25}$$

one consequence of which is that its dependence on V_α is only through the quantity $X_{\alpha\beta} = 2U_{[\alpha}V_{\beta]}$ (which reflects the fact that, apart from the normalization (3.16), ϕ does not depend on ι^A). Geometrically, this means that ϕ depends only on the point X, the conformal scale at X (through $\phi = \Omega^{-1}\phi$ under (2.16)) and the spinor o^A at X (i.e., on the twistor U_α through X). One significance of the quantity ϕ is that it is the null datum$^{(35),(36)}$ for the field $\phi_{AB\ldots L}$, defined for a null hypersurface S, where X is taken to be an arbitrary point of S and U to be the generator of S through X (see Fig. 6.6).

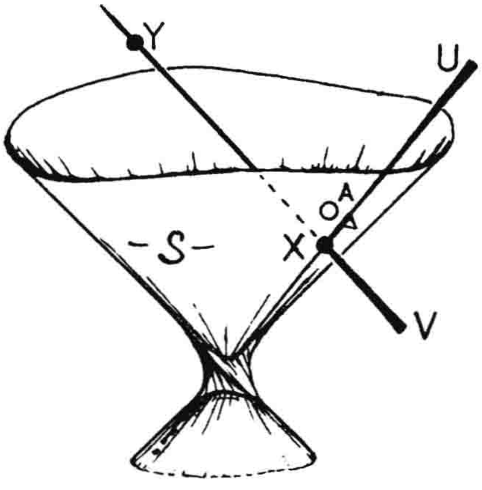


Fig. 6.6

The null datum on S can be used to define the field $\phi_{AB\ldots L}$ throughout $\mathbb{M}$ according to a certain explicit Kirchoff-type integral$^{(35),(36)}$: the field at a general point $Y \in \mathbb{M}$ is obtained by performing a 2-surface integral involving ϕ on the intersection of S with the light-cone of Y. Actually, it is a suitably defined derivative of ϕ along the generators of S which comes in, this derivative being conveniently expressed in terms of the 1-form

$$\Phi = \frac{\partial \phi}{\partial V_\alpha} \quad dU_\alpha \quad . \tag{3.26}$$

We have

$$\Phi(\nu U_\alpha, V_\alpha + \mu U_\alpha) = \Phi(U_\alpha, V_\alpha) \tag{3.27}$$

by (3.25); furthermore, by (3.22),

$$d\Phi = 0 \qquad (V_\alpha \ \underline{\text{held}}\ \underline{\text{fixed}}) \ . \tag{3.28}$$

Equation (3.28) implies that the integral

$$F(V_\alpha ; P_\alpha, Q_\alpha) = \int_{P_\alpha}^{Q_\alpha} \Phi \tag{3.29}$$

remains unaltered as the path of integration from P_α to Q_α is varied continuously in the U_α-space, provided no singularities of ϕ are crossed.

We are now in a position to state the converse of (3.10) which Bramson, Sparling and I have derived, although the full generality of the expression and a number of topological and other detailed questions remain to be sorted out. The result is that taking P_α and Q_α to be arbitrary holomorphic functions of W_β (but with suitably chosen regions for their singularities) the quantity

$$F(W_\alpha, P_\alpha(W_\beta), Q_\alpha(W_\beta)) \qquad (3.30)$$

can be used in place of $f(W_\alpha)$ in (3.10). The arbitrariness in the choice of P_α and Q_α is perhaps surprising, but it reflects the freedom (3.15). The substitution of (3.30) into (3.10), where (3.10) is evaluated at a general point Y, yields an expression which, with the situation as depicted in fig. 6.6 (for a variable point X on the light-cone of Y), can be shown to be equivalent to the Kirchoff-type integral referred to earlier. Thus we may regard the twistor function f as a sort of 'half-way' stage in the Kirchoff integral. First we have the field, defined at X, then the twistor function defined at V_α and finally the field again, defined at Y. It is worth remarking, also, that the quantity ϕ_{-1} defined by putting $r = -1$ in (3.20) (which had been considered earlier by Sparling) may be taken as (minus) an explicit integral of Φ, i.e., $-\Phi = d\phi_{-1}$ (V_α held fixed). Thus, we can write $\mathbb{F}(V_\alpha; P_\alpha, Q_\alpha) = \phi_{-1}(P_\alpha, V_\alpha) - \phi_{-1}(Q_\alpha, V_\alpha)$.

The formula (3.30) appears to have some considerable generality. In particular, for free waves on $\mathbb{M}$ of positive frequency a definite family of contours can be given, and a corresponding prescription for the functions $P_\alpha(W_\beta)$ and $Q_\alpha(W_\beta)$, so that the resulting f generates the wave globally* over $\mathbb{M}$. A similar result holds for negative frequency waves, but for a different family of contours (having the opposite twist). In each case the general location of the singularities of f is prescribed, but the singularities will not in general be poles. (Simple poles in f give rise to what are known as 'null' fields, low order poles to "algebraically special" fields.(31))

In view of the above discussion, we may take it that the natural twistor description of a massless free field is to use a one-twistor-variable holomorphic function of suitably negative homogeneity, the degree of homogeneity being related to the helicity s of the field via (3.2) and (3.4). It will be recalled, however, that the concept of helicity had already entered in a different guise in (2.8), namely as the quantity $\frac{1}{2}Z^\alpha\bar{Z}_\alpha$. Although at first sight these physical identifications seem very different, this is not really so. We may regard the discussion of §2 as referring to a single classical massless particle, whereas the field description that is being given here may be regarded as referring to the quantum description of a one-particle state.

---

* Due account must be taken, when defining fields on $\mathbb{M}$, of the phenomenon of Grgin(22), according to which a field $\phi_{A\ldots L}$ jumps by a factor i^{n+2} each time $\mathcal{I}$ is crossed. Owing to the 4-valued nature of twistors, this behaviour is taken care of automatically in (3.20).

Now, to pass from the classical to the quantum descriptions we may adopt the twistor (first-) quantization rules(32),(19)

$$[Z^{\alpha},Z^{\beta}] = 0, \quad [\bar{Z}_{\alpha},\bar{Z}_{\beta}] = 0, \quad [Z^{\alpha}, \bar{Z}_{\beta}] = \delta^{\alpha}_{\beta} \qquad (3.31)$$

which can be motivated in a number of different ways. (Units are such that $\hbar = 1$.) One such motivation is the observation that the classical relation $idZ^{\alpha} \wedge d\bar{Z}_{\alpha} = dp_{a} \wedge dx^{a}$ holds (albeit in the spinless case only) which suggests that $-iZ^{\alpha}$ and $\bar{Z}_{\alpha}$ should be regarded as canonically conjugate variables. Another motivation comes from general relativity.(32),(5) A further argument is that if the commutation rules (3.31) are adopted, then substitution into (2.11) yields momentum and angular momentum components which satisfy the standard commutation properties for generators of linear and angular displacements. This result can be obtained somewhat more rapidly from (2.3) and (2.4) by re-expressing (3.31) as

$$[\omega^{A},\omega^{B}] = 0, \quad [\omega^{A},\pi_{B'}] = 0, \quad [\pi_{A'},\pi_{B'}] = 0 \qquad (3.32)$$

and

$$[\omega^{A},\bar{\omega}^{B'}] = 0, \quad [\omega^{A},\bar{\pi}_{B}] = \varepsilon_{B}{}^{A}, \quad [\pi_{A'},\bar{\pi}_{B}] = 0 \qquad (3.33)$$

together with their conjugates. Note that there is no factor ordering problem in (2.3), (2.4) or (2.11).

On the other hand, (2.8) (and, equivalently, (2.10)) must be slightly modified when Z^α and $\bar{Z}_\alpha$ do not commute, to give the more symmetrical expression

$$s = \frac{1}{4}\ \{Z^\alpha \bar{Z}_\alpha + \bar{Z}_\alpha Z^\alpha\} \tag{3.34}$$

In fact, (3.34) follows unambiguously if we take into account the commutators (3.32) and (3.33) when (2.3) and (2.4) are substituted into (2.2) and (2.1) (for each of which there are no ambiguities in factor ordering). When considering functions of Z^α only, we can satisfy (3.31) by making the replacement

$$\bar{Z}_\alpha \mapsto -\ \partial/\partial Z^\alpha \tag{3.35}$$

Then (3.34) becomes

$$s = -\frac{1}{2}\, Z^\alpha \frac{\partial}{\partial Z^\alpha} - 1 \tag{3.36}$$

Since $Z^\alpha\partial/\partial Z^\alpha$ is the Euler homogeneity operator, its effect on the function $g(Z^\alpha)$ of (3.13) is to yield the eigenvalue $-n'-2$. Substituted into (3.36), this gives $sg = \frac{1}{2}n'g$, in agreement with (3.4). Similarly, if we are concerned with a function of W_α we can make the replacements

$$\bar{Z}_\alpha \mapsto W_\alpha, \qquad Z^\alpha \mapsto \partial/\partial W_\alpha$$

and obtain

$$s = \frac{1}{2} W_\alpha \frac{\partial}{\partial W_\alpha} + 1 \quad . \tag{3.37}$$

Operating on the $f(W_\alpha)$ of (3.9) with this we get $sf = -\frac{1}{2}nf$ from the homogeneity degree $-n-2$, of f, in agreement, now, with (3.2). Thus we have a further check on the internal consistency of twistor theory, which relates the twistor expressions for momentum and angular momentum, the twistor commutation rules and the twistor representation of free massless fields.

This is very satisfactory as far as it goes, but physical fields are not all massless, nor are massless fields always free! Furthermore, for the purposes of expressing interactions involving massless fields it might be convenient to be able, sometimes, to work in terms of potentials rather than the field quantities themselves. So long as we are restricted to twistor functions of a single variable, it seems that we are unable to represent any of these things. But with the introduction of a second twistor variable, the scheme is sufficiently widened that massive fields, massless fields with sources and potentials for massless fields can all be incorporated. (See ref. (19); also, private communication, C.J.S. Clarke.) For let $F(W_\alpha, Z^\alpha)$ be homogeneous in W_α and Z^α separately, of respective degrees $-p-2$ and $-q-2$:

$$F(\lambda W_\alpha, \mu Z^\alpha) = \lambda^{-p-2} \mu^{-q-2} F(W_\alpha, Z^\alpha) \tag{3.38}$$

where p and q are non-negative integers. We define a field

$$\Phi(x^a)_{C\ldots GM'\ldots R'} = \frac{1}{(2\pi i)^2}\oint \underbrace{W_C\ldots W_G}_{p}\ \underbrace{Z_{M'}\ldots Z_{R'}}_{q} F(W_A, -ix^{AA'}W_A, ix^{AA'}Z_{A'}, Z_{A'})\Delta WZ \tag{3.39}$$

analogously to (3.10) where, generally,

$$\Delta WX\ldots Z \ \underline{\text{means}}\ \Delta W \wedge \Delta X \wedge \ldots \wedge \Delta Z \tag{3.40}$$

(cf. (3.11)). The field has p symmetric unprimed spinor indices and q symmetric primed ones, the equation it satisfies, by virtue of its expression (3.39), now being simply

$$\nabla^{CM'}\ \Phi_{CD\ldots GM'N'\ldots R'} = 0 \tag{3.41}$$

provided pq ≠ 0. If either p or q vanishes, there are no differential constraints on the field. The contour in (3.39) need not be just a product of a contour in W_A-space with one in $Z_{A'}$-space, but is some a 2-dimensional contour in the product space. Note that (3.39) depends only on those values of F for which $W_\alpha Z^\alpha = 0$.

Equation (3.41), in its various guises, plays a number of different roles in physical theory. For example, if p = q = 1 we could use Φ_a (i.e., $\Phi_{AA'}$) as the charge-current vector in standard Maxwell theory, (3.41) then being the charge conservation law. Alternatively, we could use Φ_a for the electromagnetic potential and then (3.41) becomes the Lorenz gauge condition. For massless fields of spin greater than 1, $\Phi_{C\ldots R'}$ can still be used as a potential subject to the gauge condition (3.41)(27) Indeed, for irreducible higher spin massive fields the condition (3.41) also plays an important role. For suppose we have a spinor $\Phi_{C\ldots R'}$, symmetric in p unprimed and q primed indices which is subject to (3.41). Suppose also that

$$(\Box + m^2) \quad \Phi_{C\ldots GM'\ldots R'} = 0 \tag{3.42}$$

where $\Box = \nabla_a \nabla^a$. Then, the field

$$\Psi_{D\ldots GL'M'\ldots R'} = \nabla^{C}_{L'} \Phi_{CD\ldots GM'\ldots R'} \tag{3.43}$$

also satisfies (3.41) and (3.42) and, by virtue of (3.41), is symmetric in its q + 1 primed indices in addition to being symmetric in its p - 1 unprimed indices. By (3.42) we have, inversely,

$$\frac{m^2}{2} \Phi_{CD\ldots GM'\ldots R'} = \nabla^{L'}_{C} \Psi_{D\ldots GL'M'\ldots R'} \quad . \tag{3.44}$$

The pair of fields $(\Phi_{C\ldots R'}, 2^{\frac{1}{2}} m^{-1} \Psi_{D\ldots R'})$ related by (3.43) and (3.44) give Dirac's(25),(26) description of a particle of spin p + q. Equivalent, however, is simply to use $\Phi_{C\ldots R'}$ by itself (taking $p > 0$) subject to (3.41) and (3.42). (Dirac wanted only first order equations.) Equivalent again would be to use $\Psi_{D\ldots R'}$ alone - or, indeed, any one of the sequence of p + q + 1 fields obtainable from $\Phi_{C\ldots R'}$ by repeated application of the operators which appear on the right-hand sides of (3.43) and (3.44). The operator in (3.43) moves us one step along the sequence in the direction of increasing q and decreasing p while (apart from the factor $\frac{1}{2}m^2$) the operator in (3.44) simply moves us one step back again along the sequence. It is worth remarking, also, that if we take m = 0 we can only move outwards (in both directions) from the starting field $\Phi_{C\ldots R'}$. Coming back again we should just get zero (cf. (3.44)). In this case all the 'fields' possessing both primed and unprimed indices are considered to be merely potentials for the two at either end(27) (to ensure irreducibility). The field equations for the fields at the two ends of the sequence are simply (3.1) and (3.3) (or (3.5), if p = q = 0). See (4.35) for an illustration of this procedure (starting from one end only).

Suppose, now, that $\Phi_{C\ldots R'}$ is generated by $F(W_\alpha, Z^\alpha)$ according to (3.39), and $\Psi_{D\ldots R'}$ is related to $\Phi_{C\ldots R'}$ as in (3.43). It is easy to verify that a twistor function which generates $\Psi_{D\ldots R'}$ by the prescription (3.39) is

$$iI^{\alpha\beta}\, W_\alpha \frac{\partial F}{\partial Z^\beta} \tag{3.45}$$

In a corresponding way, if the operator of (3.44) (with a contraction on one primed index) is applied to $\Phi_{\ldots}$, the resulting field is generated by the twistor function

$$-\, iI_{\alpha\beta}\, Z^\alpha \frac{\partial F}{\partial W_\beta} \tag{3.46}$$

Application of the operator in (3.43) followed by that in (3.44) yields the operation $-\frac{1}{2}\Box$. Thus the twistor D'Alembertian may be defined as

$$\Box = -2\left(I_{\alpha\beta}\, Z^{\alpha} \frac{\partial}{\partial W_{\beta}}\right) \circ \left(I^{\rho\sigma}\, W_{\rho} \frac{\partial}{\partial Z^{\sigma}}\right) \qquad (3.47)$$

(The factors actually commute, because of (2.45).) A sufficient condition for $\Phi_{C\ldots R'}$ to satisfy (3.42) is (with the definition (3.47) for $\Box$) thus:

$$(\Box + m^2)\, F = 0 \qquad (3.48)$$

We now have a representation for massive fields in terms of two-variable twistor functions satisfying (3.48), the homogeneity degrees $-p-2$ and $-q-2$ being related to the spin j by

$$j = \frac{1}{2}(p+q) \qquad (3.49)$$

This representation does not have quite the virtue of (3.10), since in the massless case the function f was subject to no differential relation. Here we must solve (3.48). However, it should also be mentioned that there are methods whereby solutions of (3.48) may be generated starting from arbitrary two-variable twistor functions. The original prescription for doing this is due to A. Hodges$^{(37)}$. The procedure has recently been simplified by G. Sparling, the result being as follows:

$$F(W_{\alpha}, Z^{\alpha}) = \oint h(\lambda,\mu)\, G(W_{\alpha} + \lambda I_{\alpha\beta} Z^{\beta}, Z^{\alpha} + \mu I^{\alpha\beta} W_{\beta})\, d\lambda \wedge d\mu$$

where G is arbitrary (but perhaps chosen homogeneous) and h is some specific solution of

$$2\,\partial^2 h/\partial\lambda\partial\mu = m^2 h$$

(such as $\exp\{2^{-\frac{1}{2}} m(\lambda + \mu)\}$ or preferably a suitable Hankel function), the equation (3.48) then following automatically.

A natural question to ask at this juncture is whether there is anything analogous to (3.34), (3.36) and (3.37) for relating the mass and spin of a massive field to the mass and spin of a classical massive particle. Since we now have two twistor variables W_α and Z^α, we are naturally led to consider systems whose momentum and angular momentum are composed of two parts, one from each of the zero rest-mass particles described by W_α and by Z^α, respectively. In fact the momentum p_a and angular momentum M^{ab} of a general system with a future-pointing time-like momentum can be expressed (in many ways) as the sum of the momenta and of the angular momenta for two massless particles:

$$p_a = \overset{1}{p}_a + \overset{2}{p}_a \quad , \quad M^{ab} = \overset{1}{M}{}^{ab} + \overset{2}{M}{}^{ab} \tag{3.50}$$

$(\overset{1}{p}_a, \overset{1}{M}{}^{ab})$ and $(\overset{2}{p}_a, \overset{2}{M}{}^{ab})$ being each constructed from a twistor according to (2.3), (2.4) and (2.5). Call the first twistor Z^α and the second twistor $\bar{W}^\alpha$; and define the angular momentum twistor $A^{\alpha\beta}$ for the total system by

$$A^{\rho\sigma} \longleftrightarrow \begin{pmatrix} -2i\mu^{RS} & p^{R}{}_{S'} \\ & \\ p_{R'}{}^{S} & 0 \end{pmatrix} \tag{3.51}$$

where

$$M^{AA'BB'} = \mu^{AB}\varepsilon^{A'B'} + \varepsilon^{AB}\bar{\mu}^{A'B'} \quad \text{and} \quad \mu^{AB} = \mu^{BA} \ . \tag{3.52}$$

Then

$$A^{\alpha\beta} = 2Z^{(\alpha}I^{\beta)\gamma}\,\bar{Z}_{\gamma} + 2\bar{W}^{(\alpha}I^{\beta)\gamma}\,W_{\gamma} \tag{3.53}$$

the two terms on the right being the angular momentum twistors for the two constituent massless particles. (A similar expression would hold for a system composed of any finite number of parts).

Because of the skew-symmetry of $I^{\alpha\beta}$ there is no factor ordering problem in (3.53), where we have (taking the 'quantum' version of (3.53))

$$[Z^{\alpha}, \bar{Z}_{\beta}] = \delta^{\alpha}_{\beta} \ , \quad [W_{\alpha}, \bar{W}^{\beta}] = -\,\delta^{\beta}_{\alpha} \tag{3.54}$$

all other commutators vanishing (cf. (3.31)). The commutation relations for momentum and angular momentum can be expressed in twistor form as

$$\begin{aligned} [A^{\alpha\beta}, A^{\rho\sigma}] &= 2A^{\alpha(\rho}I^{\sigma)\beta} + 2A^{\beta(\rho}I^{\sigma)\alpha} \ , \\ [A^{\alpha\beta}, \bar{A}_{\rho\sigma}] &= 4\delta^{(\alpha}_{(\rho}I_{\sigma)\gamma}\,A^{\beta)\gamma} = 4\delta^{(\alpha}_{(\rho}\,I^{\beta)\gamma}\,\bar{A}_{\sigma)\gamma} \end{aligned} \tag{3.55}$$

and this follows directly from (3.53) and (3.54). (In terms of $p_{AA'}$ and μ^{AB} these relations are $[p_{AA'}, p_{BB'}] = 0$, $[p_{AA'}, \mu^{BC}] = i\varepsilon_{A}{}^{(B} p_{A'}{}^{C)}$, $[\mu^{AB}, \mu^{CD}] = i\mu^{A(C}\varepsilon^{D)B} + i\mu^{B(C}\varepsilon^{D)A}$, $[\mu^{AB}, \bar{\mu}^{C'D'}] = 0$.)

The two expressions for the second commutator in (3.55) are equal because the angular momentum twistor possesses the symmetries

$$I_{\alpha\gamma} A^{\beta\gamma} = \bar{A}_{\alpha\gamma} I^{\beta\gamma}, \quad A^{\alpha\beta} = A^{\beta\alpha} \tag{3.56}$$

These follow from the fact that $A^{\alpha\beta}$ has the form

$$A^{\alpha\beta} = 2E_{\gamma}^{(\alpha} I^{\beta)\gamma} \tag{3.57}$$

where E^{α}_{β} is Hermitian:

$$\bar{E}^{\alpha}_{\beta} = E^{\alpha}_{\beta} \tag{3.58}$$

Here we can take

$$E^{\alpha}_{\beta} = Z^{\alpha}\bar{Z}_{\beta} + W_{\beta}\bar{W}^{\alpha} = \bar{Z}_{\beta}Z^{\alpha} + \bar{W}^{\alpha}W_{\beta} \tag{3.59}$$

(cf. (3.54)) and (3.53) will be satisfied. Since the addition of a trace term to E^{α}_{β} makes no difference to (3.57) (because of the skew-symmetry of $I^{\alpha\beta}$) other relative factor orderings for the two terms in (3.59), although less symmetrical, could also be used. Alternatively, the trace of E^{α}_{β} could be removed altogether and the quantity

$$E^{\alpha}_{\beta} - \frac{1}{4}\delta^{\alpha}_{\beta} E^{\gamma}_{\gamma} \tag{3.60}$$

used in place of E^{α}_{β} in (3.57). The significance of the particular expression (3.60) is that its 15 independent components generate the Lie algebra of SU(2.2) - i.e., of the conformal group - in the same way that the 10 independent components of $A^{\alpha\beta}$ (and $\bar{A}_{\alpha\beta}$) generate the Lie algebra of the Poincare group. The commutation relations for E^{α}_{β} (by virtue of (3.54)) are

$$[E^{\alpha}_{\beta}, E^{\rho}_{\sigma}] = \delta^{\alpha}_{\sigma} E^{\rho}_{\beta} - \delta^{\rho}_{\beta} E^{\alpha}_{\sigma} \tag{3.61}$$

(giving the Lie algebra of U(2,2)). The commutation relations for (3.60) are then obtained by simply replacing each E^{α}_{β} in (3.61) by $E^{\alpha}_{\beta} - \frac{1}{4}\delta^{\alpha}_{\beta} E^{\gamma}_{\gamma}$. The commutation relations for $A^{\alpha\beta}$, $\bar{A}_{\alpha\beta}$ follow also from these.

The operators for squared mass $m^2 = p_a p^a$ and squared total spin $- S_a S^a m^{-2}$ (S_a being the Pauli-Lubanski vector (2.2), but now in the massive case) can now be calculated. We find(19)

$$m^2 = 2(I_{\alpha\beta} Z^{\alpha} \bar{W}^{\beta}) (I^{\rho\sigma} \bar{Z}_{\rho} W_{\sigma}) \tag{3.62}$$

and*

---

*I am indebted to P. Tod for obtaining this formula for me (and also to Z. Perjés for having earlier found the classical version of it).

$$- S_a S^a m^{-2} = j(j + 1) + W_\alpha Z^\alpha \bar{W}^\beta \bar{Z}_\beta \tag{3.63}$$

where

$$2j = Z^\alpha \bar{Z}_\alpha - \bar{W}^\alpha W_\alpha = \bar{Z}_\alpha Z^\alpha - W_\alpha \bar{W}^\alpha = Z^\alpha \bar{Z}_\alpha - W_\alpha \bar{W}^\alpha - 4. \tag{3.64}$$

The function $F(W_\alpha, Z^\alpha)$ that we have been considering is, indeed, an eigenstate of the operator (3.62) and, modulo functions of the form $W_\alpha Z^\alpha h$, also of the operator (3.63). (Recall that (3.39) refers only to the region of (W_α, Z^α)-space where $W_\alpha Z^\alpha = 0$.) This follows from (3.48), with (3.47), and (3.38) (where $\bar{W}^\alpha \to \partial/\partial W_\alpha$ and $\bar{Z}_\alpha \to - \partial/\partial Z^\alpha$). The eigenvalue in the case (3.62) is the correct squared mass, in agreement with (3.48). In the case of (3.63) we obtain the correct quantum mechanical expression $j(j + 1)$, the eigenvalue of the spin operator j being $\frac{1}{2}(p + q)$, in agreement with (3.49). This follows from $Z^\alpha \bar{Z}_\alpha F = (p + 2)F$ and $W_\alpha \bar{W}^\alpha F = -(q + 2)F$ because of the homogeneity degrees of F. Thus we have another successful check on the consistency of the twistor quantization rules. (Note, also, that twistor theory has an advantage over conventional theory in having a simple expression for the operator j.) It is possible to go further in the interpretation of the twistor operators involved in E^α_β, but the matter will not be persued here.

The attitude adopted here has been to regard massive systems to be, in a certain sense, composed of two (or perhaps more) massless systems. However, as was remarked briefly earlier, the decomposition into two massless systems is very far from unique. It is of some interest to investigate this non-uniqueness since we could regard the group of transformations between Z^α, $\bar{Z}_\alpha$, W_α, $\bar{W}^\alpha$ which leaves $A^{\alpha\beta}$ invariant as a kind of internal symmetry group for a massive particle. It is not hard to show, in fact, that the group which preserves the form (3.59) for E^α_β (where we may regard this now as just a classical expression) is simply U(2). The group preserving $A^{\alpha\beta}$ is just slightly bigger and incorporates transformations of the type

$$Z^\alpha \mapsto Z^\alpha + \lambda I^{\alpha\beta} W_\beta, \quad W_\alpha \mapsto W_\alpha - \bar{\lambda} I_{\alpha\beta} Z^\beta$$

together with their conjugates. The group turns out to be

$$SU(2) \times \tilde{E}(2) \quad ,$$

$\tilde{E}(2)$ being a twofold covering of the group of Euclidean motions of the plane.*

There is much else that can be said concerning the angular momentum twistor $A^{\alpha\beta}$, for example its relation to Coulomb-type fields

---

* The corresponding group preserving an $A^{\alpha\beta}$ composed of three twistors and their conjugates, rather than two, has been found by Z. Perjés(50) to be isomorphic (at least locally) with the 15-parameter 'inhomogeneous U(3)'. He has also found that its 14-parameter 'inhomogeneous SU(3)' subgroup acts (simply) transitively on the space of triplets of twistors which compose $A^{\alpha\beta}$.

and to their spinning analogues, the Kerr-type fields. The reader is referred to ref. (19) for a brief discussion.

4. TWISTOR DIAGRAMS

One of the main aims of twistor theory is to provide an alternative approach to the quantization of fields. As yet, no complete prescription has emerged, partial results having been obtained mainly by the direct translation of Feynman (position space) graphs for massless particles and of other important expressions, such as charge integrals etc. It seems that the resulting expressions can normally be diagrammatically represented according to certain well-defined rules and although these rules must be regarded as somewhat provisional, they seem to point to the emergence of some new kinds of insight as to the possible nature of interactions between particles. Also, it might eventually, turn out that the twistor diagrams - or some modification of them - can provide the sought-for generalization of the spin-networks that were mentioned in §1.

I shall state these tentative rules for twistor diagrams first and then show how a few simple processes can be described in terms of such diagrams. The diagrams themselves consist of a number of black and white[*] spots called vertices connected by line segments (simple or multiple) called edges where each edge joins a black to a white vertex (internal edge) - or else has a free end (external edge). Every edge bears a number - usually an integer. The free ends are normally labelled by letters, and they represent fixed twistors. The vertices themselves represent variable twistors which are to be integrated over, black vertices denoting $\begin{bmatrix}1\\0\end{bmatrix}$ -twistors Z^α, Y^α, ... and white vertices, $\begin{bmatrix}0\\1\end{bmatrix}$ -twistors W_α, V_α, ...

---

[*] Since when writing on a blackboard, the roles of the colours are reversed, it is sometimes convenient to refer to these vertices as "filled" and "open", respectively.

The whole diagram defines a closed differential form, and is intended also to give some indication of the contour over which this form is to be integrated. The differentials involved* are

$$
\begin{aligned}
\mathbb{D}\,ZY\ldots WV\ldots &= d^4Z \wedge d^4Y \wedge \ldots \wedge d^4W \wedge d^4V \wedge \ldots \\
&= dZ^0 \wedge dZ^1 \wedge dZ^2 \wedge dZ^3 \wedge dY^0 \wedge \ldots \wedge dW_0 \wedge \ldots \wedge dW_3 \wedge dV_0 \wedge \ldots \\
&= (\tfrac{1}{24}\, \varepsilon_{\alpha\beta\gamma\delta}\, dZ^\alpha \wedge dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta) \wedge (\ldots) \wedge \ldots \qquad (4.1)
\end{aligned}
$$

corresponding to all the vertices, where (4.1) is multiplied by an additional factor corresponding to each edge. To determine this factor, we add the number appearing on the edge to the multiplicity k of the edge - call the result r - then if r is a positive integer, the factor is

---

* Slightly different differential forms were employed in refs. (19), (32). Their relation to (4.1) is discussed in (4.8), (4.9) and (4.10).

$$\frac{(-1)^{r-1}\,(r-1)!}{(2\pi i)^{k+1}\,(W_\alpha Z^\alpha)^r} \tag{4.2}$$

where W_α and Z^α are the twistors represented by the ends of the edge. Further, if r is any complex number except zero or a negative integer then the generalization

$$(2\pi i)^{-k}\,(W_\alpha Z^\alpha)_r \tag{4.3}$$

is employed, where

$$(x)_r = -\frac{1}{2\pi i}\,\Gamma(r)\,(-x)^{-r} \tag{4.4}$$

Note that (4.4) satisfies

$$\frac{d(x)_r}{dx} = (x)_{r+1} \tag{4.5}$$

When r is a non-positive integer the factor is to be

$$\frac{(W_\alpha Z^\alpha)^{-r}}{(2\pi i)^k\,(-r)!} \tag{4.6}$$

but there is also a crucial difference in the nature of the contour that is to be used whenever (4.6) is employed. This point will be elaborated shortly. There are also other ways of handling the cases when r is a non-positive integer, such as applying a limiting process to (4.3) or using logarithmic terms. These are essentially the methods that had been used in earlier work(19),(32) but they will not be considered here. The key point is that one should be able to integrate by parts consistently in a process which reduces the value r from 1 to 0. This will be considered in a little detail shortly for the expression (4.6). The definition of the differential form represented by a twistor diagram is now completed by incorporating one factor* of $2\pi i$ in the numerator for each free end.

There are certain (provisional) rules that will usually be adopted in order that the diagrams will be considered as 'correctly drawn'. These are, first, that the sum of the numbers on all edges having a common vertex should be zero. Secondly, there should be exactly four edges meeting at each vertex, counting multiple edges according to their multiplicity. (Note that this implies that no edge can be more than quadruple.) Occasionally this rule can be relaxed (but normally only for edges for which r is a non-positive integer) in which case the first two rules are replaced by the statement that the sum of the 'r-values' of the edges at each vertex must be 4. Finally there are the rules that all the external edges should be simple and that

---

* The factors of $2\pi i$ that appear at various places are chosen so that the answers to simple diagrams are comparatively neat expressions. There is little in the way of a clear-cut theoretical justification of these factors.

(normally) no two vertices may be joined by distinct edges.

These are the rules in the case of explicit conformally invariant diagrams. There is another type of factor which can come in when conformal invariance is not required, namely

$$Z^{\alpha}Y^{\beta}I_{\alpha\beta} \quad \underline{\text{or}} \quad W_{\alpha}V_{\beta}I^{\alpha\beta} \tag{4.7}$$

Such factors (in the numerator) will be denoted by dotted lines joining the vertices in question, both of which must be of the same colour (i.e., both black for the first expression (4.7) and both white for the second). For a squared factor (4.7) in the numerator, two distinct dotted lines connecting the same pair of vertices may be used; for a cubed factor, three distinct dotted lines, etc. One can also introduce pole terms analogous to (4.2) or branching singularities analogous to (4.3). Such expressions are needed for Hodges' formula for massive fields but will not be entered into here. Concerning the rule that there should be four lines (counting multiplicities) at each vertex, dotted lines do not count at all. (The rule has to do with the topology of the singularity sets and the factors (4.7) are non-singular.) However, the rule for the number count at each vertex must be slightly amended in that each dotted line at the vertex contributes - 1 to the sum at that vertex. (This rule is imposed in order to give a homogeneity degree of - 4 in the variable twistor representing the vertex, the homogeneity degree in (4.7) being + 1.) I shall also allow dotted lines joining vertices of opposite colour to denote scalar product factors $W_\alpha Z^\alpha$ in the numerator and which (in contradistinction to (4.6)) does not enter into the contour location. Figure 6.7 summarizes the basic rules for the diagrams.

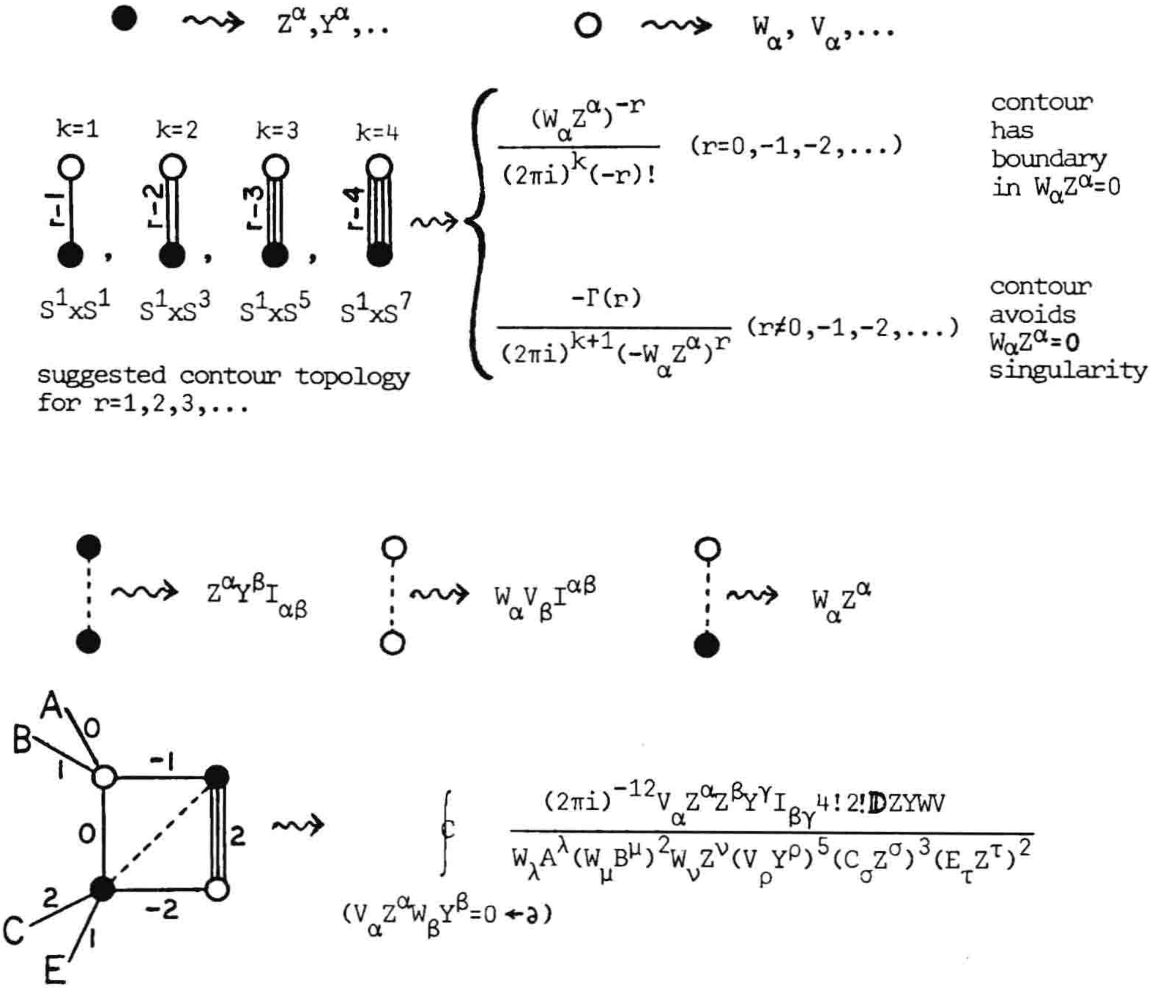

$\rightsquigarrow Z^\alpha, Y^\alpha, \ldots$
$\rightsquigarrow W_\alpha, V_\alpha, \ldots$
k=1
k=2
k=3
k=4
r-1
r-2
r-3
r-4
$S^1 \times S^1$
$S^1 \times S^3$
$S^1 \times S^5$
$S^1 \times S^7$
suggested contour topology for r=1,2,3,...
$\dfrac{(W_\alpha Z^\alpha)^{-r}}{(2\pi i)^k(-r)!}$ (r=0,-1,-2,...)
contour has boundary in $W_\alpha Z^\alpha=0$
$\dfrac{-\Gamma(r)}{(2\pi i)^{k+1}(-W_\alpha Z^\alpha)^r}$ ($r\neq 0,-1,-2,\ldots$)
contour avoids $W_\alpha Z^\alpha = 0$ singularity
$\rightsquigarrow Z^\alpha Y^\beta I_{\alpha\beta}$
$\rightsquigarrow W_\alpha V_\beta I^{\alpha\beta}$
$\rightsquigarrow W_\alpha Z^\alpha$
A
B
C
E
0
1
-1
0
2
-2
2
1
$\rightsquigarrow \oint \dfrac{(2\pi i)^{-12} V_\alpha Z^\alpha Z^\beta Y^\gamma I_{\beta\gamma} 4!\,2!\,\mathbb{D}ZYWV}{W_\lambda A^\lambda (W_\mu B^\mu)^2 W_\nu Z^\nu (V_\rho Y^\rho)^5 (C_\sigma Z^\sigma)^3 (E_\tau Z^\tau)^2}$
$(V_\alpha Z^\alpha W_\beta Y^\beta = 0 \leftarrow \partial)$

Fig. 6.7

We now come to the question of the contour over which this differential form is to be integrated. Provided the 'r-values' corresponding to all the edges are positive (or non-integral) the contour is to be some closed (i.e., compact without boundary) 4v-dimensional surface in the product space of all the twistors Z^α, $Y^\alpha, \ldots, W_\alpha, V_\alpha, \ldots$ which correspond to the vertices, these being v in number. Unfortunately there is as yet no clear topological rule as to how this contour is to be chosen. In an appropriate sense - as yet to be formulated in some precise way - one may require the contour to surround the singularity corresponding to each edge, where $r = 1,2,3,\ldots,$ so that for a simple edge it surrounds it with a torus $S^1 \times S^1$, for a double edge, an $S^1 \times S^3$, a triple edge, $S^1 \times S^5$ and for a quadruple edge, $S^1 \times S^7$. However, even if these prescriptions could be formulated precisely, some (discrete) ambiguities would remain in the choice of contour for many diagrams. Also, it seems likely that for certain diagrams these prescriptions cannot be satisfied at all. The topological problems can become quite complicated and seem to require a considerable use of the machinery of algebraic topology. Several results have been obtained in this way by Sparling and are described at length in his article in this volume(38). But there appear to be many further insights yet to be gained along these lines.

For edges with non-integral r, the contours, while still closed, are different from those with integral r owing to the many-sheeted Riemann surfaces involved. Some results for these cases have also been obtained by Sparling and are described in his Ph.D. thesis(39). Such twistor diagrams will not be considered here in detail, however.

When the diagram contains edges for which $r = 0, -1, -2, \ldots,$ the rule (following a suggestion(39) of Sparling's) is that the contour is *not* required to be closed but must have *its boundary in the locus consisting of the union of all regions defined by the vanishing of the twistor scalar products associated with those particular edges*. Since when $r = 0, -1, -2, \ldots$ the expression (4.6) is non-singular, this means that again we are integrating expressions which are finite throughout the entire compact region of integration. In all cases the poles $W_\alpha Z^\alpha = 0$ in (4.2) must be avoided by the contour (as must the branching singularities in (4.3)). Thus, in all cases, the results of the integrals defined by twistor diagrams must necessarily be *finite* (or zero). It should also be emphasized that all the integrals so performed, whether they are over bounded contours or closed ones, are genuine contour integrals. That is to say, if the region of integration is deformed continuously, while continuing to satisfy the conditions specified above, then the result of the integration will be unaffected. This follows from the properties of differential forms and the holomorphic nature the integrand.

There is a point of difference between the interpretation of the twistor diagrams as presented here and that which had been given earlier(19),(32). This concerns the use of the differential forms (4.1) rather than

$$\mathcal{D}ZY\ldots WV\ldots = \mathcal{D}Z \wedge \mathcal{D}Y \wedge \ldots \wedge \mathcal{D}W \wedge \ldots \qquad (4.8)$$

where

$$\mathcal{D}Z = \frac{1}{6}\,\varepsilon_{\alpha\beta\gamma\delta}\, Z^{\alpha}\, dZ^{\beta} \wedge dZ^{\gamma} \wedge dZ^{\delta}$$

$$= Z^{0}\, dZ^{1} \wedge dZ^{2} \wedge dZ^{3} - Z^{1}\, dZ^{0} \wedge dZ^{2} \wedge dZ^{3} \qquad (4.9)$$

$$+ Z^{2}\, dZ^{0} \wedge dZ^{1} \wedge dZ^{3} - Z^{3}\, dZ^{0} \wedge dZ^{1} \wedge dZ^{2}$$

the above forms being appropriate for integrations within a projective 3-space. To see the equivalence of the use of (4.1) with (4.8) consider the integrations of twistor variables one at a time. We have

$$\oint f(Z^{\alpha})\, d^{4}Z = \oint (Z^{0})^{4}\, f(Z^{\alpha})\, \frac{dZ^{0}}{Z^{0}} \wedge d\left(\frac{Z^{1}}{Z^{0}}\right) \wedge d\left(\frac{Z^{2}}{Z^{0}}\right) \wedge d\left(\frac{Z^{3}}{Z^{0}}\right)$$

$$= \oint \left[\sum_{r=-\infty}^{\infty} (Z^{0})^{4+r}\, f_{r}\left(\frac{Z^{\alpha}}{Z^{0}}\right)\right] \frac{dZ^{0}}{Z^{0}} \wedge d\left(\frac{Z^{1}}{Z^{0}}\right) \wedge d\left(\frac{Z^{2}}{Z^{0}}\right) \wedge d\left(\frac{Z^{3}}{Z^{0}}\right)$$

$$= 2\pi i \oint f_{-4}(Z^{\alpha})\, (Z^{0})^{4}\, d\left(\frac{Z^{1}}{Z^{0}}\right) \wedge d\left(\frac{Z^{2}}{Z^{0}}\right) \wedge d\left(\frac{Z^{3}}{Z^{0}}\right) \qquad (4.10)$$

$$= 2\pi i \oint f_{-4}(Z^{\alpha})\quad \mathcal{D}Z$$

where the Z^{0}-contour is suitably chosen and where $f = \Sigma f_{r}$ is some expansion of f into parts homogeneous of each degree r. Thus, for a function which is already of homogeneity degree -4, it makes no difference, apart from the factor $2\pi i$, whether the form $d^{4}Z$ or $\mathcal{D}Z$ is used. Furthermore, the rules for twistor diagrams ensure that the integrands are indeed homogeneous of degree -4 in each variable.

But the use of d^4Z seems to offer more possible scope, since f need not be homogeneous. The integration picks out a 'part' of f which is homogeneous of degree -4.

In an exactly similar way, the forms (3.11) may be replaced by

$$d^2W = dW_0 \wedge dW_1 = \frac{1}{2}\varepsilon^{AB}\,dW_A \wedge dW_B = \frac{1}{2} I^{\alpha\beta}\,dW_\alpha \wedge dW_\beta$$

$$d^2Z = dZ^0 \wedge dZ^1 = \frac{1}{2}\varepsilon^{A'B'}\,dZ_{A'} \wedge dZ_{B'} = \frac{1}{2} I_{\alpha\beta}\,dZ^\alpha \wedge dZ^\beta \quad . \qquad (4.11)$$

Thus (3.10) can be re-expressed as

$$\phi_{AB\ldots L} = \frac{1}{(2\pi i)^2}\oint W_A\,W_B\,\ldots\,W_L\,f(W_P,\,-ix^{PP'}\,W_P)\,d^2W \qquad (4.12)$$

and the zero rest-mass equations (3.1) are automatically satisfied, as before, but where now we do not need to assume that $f(W_\alpha)$ is homogeneous of degree $-n-2$. If f is not so homogeneous, (4.12) picks out a part which is.

Let us next consider how the twistor diagrams are to be employed in a physical context. The idea is that a scattering amplitude is to be calculated, where the states of the incoming and outgoing particles are described by means of twistor holomorphic functions as in (3.10) (cf. (4.12)) or (3.12), in the case of massless particles, or by use of (3.39) in the case of massive particles. I shall consider massless particles only here. (See ref. (37) for a discussion of the massive case.) The in-states and out-states are then to be suitable wave packets described by holomorphic functions of single twistor variables.

There are certain specific types of holomorphic function, namely

$$f(W_\alpha) = \frac{Q}{(W_\alpha A^\alpha)^{p+1}\,(W_\beta B^\beta)^{q+1}} \quad \underline{\text{and}} \quad g(Z^\alpha) = \frac{R}{(C_\alpha Z^\alpha)^{r+1}(D_\beta Z^\beta)^{s+1}} \tag{4.13}$$

which fit especially well into the framework of the twistor diagrams. If the join of the points A, B in $\mathbb{PT}$ and the intersection of the planes C, D in $\mathbb{PT}$ give lines which do not meet $\mathbb{N}$, i.e., which correspond to a point in $\mathbb{M}$ with a timelike imaginary part, then the wave-packets are non-singular. Such wave packets are called elementary states. If the line lies entirely in $\mathbb{PT}^-$ then the wave packet has positive frequency; if it lies entirely in $\mathbb{PT}^+$, the wave has negative frequency(32). Elementary states form an overcomplete set for each spin.

It will be noticed that the expressions (4.13) are (apart from a possible constant factor) just of the type that arise from a pair of external edges joined to one vertex of a twistor diagram (cf. Fig. 6.7). In fact it will turn out that twistor diagrams, in which all external edges meet in pairs at vertices of the diagram, can often represent amplitudes (or partial amplitudes) for processes in which all external states are elementary states. We can, indeed, also represent any in- and out-states as part of a diagram, whether they are elementary states or not, but then the diagrammatic description may not be explicit. We shall see some examples of this in a moment.

By way of introduction to the techniques for translating expressions into twistor diagrams, I shall first consider the charge integral for an electromagnetic field - even though this expression is not strictly a twistor diagram. Suppose that ϕ_{AB} describes an electromagnetic field, according to (3.7), which is source-free outside a certain timelike world-tube. In this exterior region the field ϕ_{AB} will satisfy (3.1) and will be assumed representable in the form (3.10) by means of a holomorphic twistor function $f(W_\alpha)$ which is homogeneous of degree - 4. Now the total charge e of the source is given by

$$e = \oint \overset{*}{F}_{ab}\, dx^a \wedge dx^b = -2i \oint \phi_{AB}\, dx^A_{A'} \wedge dx^{BA'} \qquad (4.14)$$

integrated over some closed 2-surface which surrounds it. Substituting (4.12) into (4.14) we get

$$e = \frac{i}{2\pi^2} \oint f(W_\alpha)\, W_A W_B\, d^2W \wedge dx^A_{A'} \wedge dx^{BA'} \quad . \qquad (4.15)$$

But

$$W_A W_B\, d^2W \wedge dx^A_{A'} \wedge dx^{BA'} = 2dW_o \wedge dW_1 \wedge d(W_A x^{AO'}) \wedge d(W_B x^{B1'})$$

$$= -2d^4W \qquad (4.16)$$

since W_α passes through x^a; i.e., $W_\alpha \leftrightarrow (W_A, -ix^{AA'}W_A)$.

Substituting into (4.15) we get

$$e = \frac{1}{\pi^2 i} \oint f(W_\alpha)\, d^4W = \frac{2}{\pi} \oint f(W_\alpha)\ \mathcal{D}W \ . \tag{4.17}$$

In a similar way, if $g(W_\alpha)$ is homogeneous of degree -6 and describes a linearized gravitational field according to (4.12), (3.8), which is source-free outside a timelike world-tube,

$$\bar{A}_{\alpha\beta} = h \oint W_\alpha W_\beta\ g(W_\gamma)\ d^4W \tag{4.18}$$

(where h is some constant) defines the energy-momentum and angular momentum of the source according to (3.51).

The total charge of a system can also be calculated as the 3-surface integral

$$e = \int j^a\, d^3x_a \tag{4.19}$$

over a suitable hypersurface S, where j^a is the charge-current vector, satisfying

$$\nabla_a\, j^a = 0 \tag{4.20}$$

and where

$$d^3x_a = \frac{1}{3!}\ \varepsilon_{abcd}\ dx^b \wedge dx^c \wedge dx^d \tag{4.21}$$

gives the 3-volume element of S. Because of (4.20) (cf. (3.41)) we can reasonably expect to generate j_a from a two-variable twistor function $J(W_\alpha, Z^\alpha)$, homogeneous of degrees $(-3,-3)$, according to (3.39), so that (4.19) becomes

$$e = \frac{1}{(2\pi i)^4} \oint J(W_\alpha, Z^\alpha)\, W_A Z_{A'}\, d^2W \wedge d^2Z \wedge d^3x^{AA'} \tag{4.22}$$

where W_α and Z^α both pass through x^a. To express (4.22) as an integral over the two twistors W_α and Z^α varying independently, we must introduce what is, in effect, a δ-function of $W_\alpha Z^\alpha$ since $W_\alpha Z^\alpha = 0$ is the condition that W_α and Z^α do in fact pass through the same point.* With contour integrations of holomorphic functions we can use Cauchy's $f(0) = (2\pi i)^{-1} \oint f(z)\, z^{-1}\, dz$, rather than Dirac's $f(0) = \int f(x)\delta(x)dx$. Thus we pull W_α and Z^α into incidence with one another by introducing

$$\frac{d(W_\alpha Z^\alpha)}{2\pi i\; W_\beta Z^\beta} \tag{4.23}$$

into the integral. Now, we have, in effect,

---

* Since this is a holomorphic restriction on W_α and Z^α whereas, for example, the condition that the point x^a be real (or, say, that W_α be null) is no restriction holomorphically, it is the former that must be specified explicitly in some way in the integral, while the latter is merely a statement concerning the location of the contour.

$$d^2W \wedge d^2Z \wedge W_A Z_{A'},\ d^3x^{AA'} \wedge d(W_\alpha Z^\alpha) = 2d^4W \wedge d^4Z \qquad (4.24)$$

although for this formula to be an identity, with

$$W_\alpha \leftrightarrow (W_A, -ix^{AA'}W_A),\quad Z^\alpha \leftrightarrow (iy^{AA'}Z_{A'}, Z_{A'}) \qquad (4.25)$$

some further specification of the relation between x^a and y^a is required. We can achieve this by making a choice of x^a and y^a for each general W_α and Z^α such that W_α always passes through x^a and Z^α always passes through y^a, where $x^a = y^a$ whenever $W_\alpha Z^\alpha = 0$. This can be done in many ways. Indeed, it is sufficient to verify (4.24) in the very special case where S is given by $x^0 = 0$, i.e., by $x^{00'} + x^{11'} = 0$. Then we can choose to relate y^a to x^a by

$$y^{00'} = -x^{11'},\quad y^{11'} = -x^{00'},\quad y^{01'} = x^{01'},\quad y^{10'} = x^{10'}$$

The requirements stated above (including (4.24)) can now be easily verified directly*. Thus, incorporating (4.23) into (4.22) and using

---

* Somewhat more generally we can select a non-singular skew-symmetric twistor $S^{\alpha\beta}$ with inverse $T_{\alpha\beta}$ and let y^a be defined (as in (2.33)) by $Y^{\alpha\beta} = Z^{[\alpha}S^{\beta]\gamma}W_\gamma$ and x^a by $X_{\alpha\beta} = W_{[\alpha}T_{\beta]\gamma}Z^\gamma$. The points for which $x^a = y^a$ then lie on a complex hypersurface which (owing to the surface-independence of (4.19)) can be taken to be S. This surface is here given by $X_{\alpha\beta}S^{\alpha\beta} = 0$ (which defines a system of lines in $\mathbb{PT}$ known as a linear complex$^{(24)}$. We can specify y^a as a function of x^a by $Y^{\alpha\beta} = S^{\alpha\gamma}X_{\gamma\delta}S^{\delta\beta}$ and then (4.24) turns out to be true.

(4.24) we get the result

$$e = \frac{2}{(2\pi i)^5} \oint \frac{J(W_\beta, Z^\beta)}{W_\alpha Z^\alpha} \, d^4W \wedge d^4Z \qquad (4.26)$$

This formula will be related to (4.17) in due course.

A special case of (4.26) gives us our first twistor diagram, namely that for the scalar product between two massless spin $\frac{1}{2}$ fields. For let

$$j_{AA'} = \phi_A \psi_{A'} \qquad (4.27)$$

where ϕ_A and $\psi_{A'}$ are solutions of the Weyl neutrino equation ((3.1) and (3.3) with $n = 1 = n'$) and so can be given in terms of twistor functions $f(W_\alpha), g(Z^\alpha)$, each homogeneous of degree -3, by

$$\phi(x^p)_A = \frac{1}{(2\pi i)^2} \oint W_A \, f(W_\alpha) \, d^2W, \qquad \psi(x^p)_{A'} = \frac{1}{(2\pi i)^2} \oint Z_{A'} \, g(Z^\alpha) \, d^2W. \qquad (4.28)$$

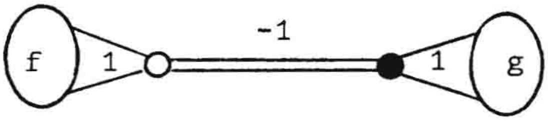
The scalar product between $\bar{\psi}_A$ and ϕ_A is given(27) by

$$\langle \bar{\psi} | \phi \rangle = k \oint \phi_A \, \psi_{A'} \, d^3 x^{AA'} \qquad (4.29)$$

(for a suitable constant k) which is surface-independent because $\nabla^{AA'}(\phi_A \psi_{A'}) = 0$ (cf. (4.20)) the 3-surface *S* being now closed with the point I at infinity. Substituting (4.27) into (4.29) and comparing with (4.19), (4.26) we get

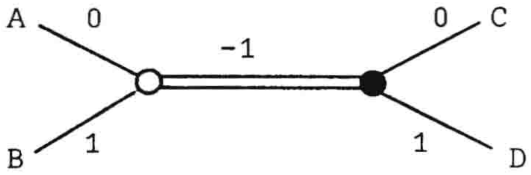
$$\langle\bar{\psi}|\phi\rangle = \frac{2k}{(2\pi i)^5}\oint \frac{f(W_\beta)\ g(Z^\gamma)}{W_\alpha Z^\alpha}\ d^4W \wedge d^4Z\ . \qquad (4.30)$$

The diagram for (4.30) (ignoring the overall factor) is



(4.31)

which, in the special case when f and g generate elementary states given by (4.13) with p = 0, q = 1; r = 0, s = 1, becomes



(4.32)

In (4.31) the encircled f and g simply represent the functions themselves, the two lines emanating from each circle indicating that the singularity structure is to be similar to that of the corresponding elementary states, as denoted explicitly in (4.32). The two numbers '1' appearing in (4.31) indicate simply that the homogeneity degree of each of f and g is − 3, the '1' and each line contributing a count − 1 to the homogeneity. (It is not meaningful, in the case of general functions, to attribute a homogeneity to each singularity separately.) The double line in the centre of (4.32) (and (4.31)) is intended to indicate the existence of an S^3 contribution to the shape of the contour, in accordance with the scheme depicted in Fig. 6.7. In fact the entire contour for (4.32) can indeed be taken to have the structure $S^1 \times S^1 \times S^1 \times S^1 \times (S^1 \times S^3)$. (See Sparling's article in this volume for a discussion of the topology of relevance to (4.32).)

The scalar product for massless fields of spin zero is

$$\langle\bar{\psi}|\phi\rangle = -\,ik \oint \frac{1}{2}\{\phi\nabla_a\psi - \psi\nabla_a\phi\}\, d^3x^a \qquad (4.33)$$

while that for higher spin $\frac{1}{2}n$ can be expressed as

$$\langle\bar{\psi}|\phi\rangle = i^{n-1}\, k \oint \underbrace{\phi_{AB\ldots L}}_{n}\ \overset{(1)}{\psi}{}_{A'}{}^{\overbrace{B\ldots L}^{n-1}}\ d^3x^{AA'} \quad , \qquad (4.34)$$

where $\overset{(1)}{\psi}{}_{A'}^{B\ldots L}$ is a potential for the field $\psi_{A'B'\ldots L'}$ related to it via the chain$^{(27)}$

$$\nabla^{FF'}\ \overset{(r)}{\psi}{}^{G\ldots L}_{A'\ldots E'F'} = 0, \quad \nabla_{GG'}\ \overset{(r)}{\psi}{}^{GH\ldots L}_{A'\ldots F'} = \overset{(r+1)}{\psi}{}^{H\ldots L}_{A'\ldots F'G'}\ ;$$

$$\overset{(n)}{\psi}{}_{A'\ldots L'} = \psi_{A'\ldots L'} \qquad (4.35)$$

It may be verified that the expressions (4.33), (4.34) are 3-surface independent, Hermitian symmetric and gauge invariant. That they are also conformally invariant will follow from their twistor translations (4.49) (cf. also ref. (40)).

The expressions (4.33), (4.29) and (4.34) can be related to one another by employing the observation that if Q^A satisfies the twistor equation (2.13)

$$\nabla_{B'}^{(B}\, Q^{A)} = 0 \qquad (4.36)$$

then

$$\gamma_{A\ldots K} = \phi_{A\ldots KL}\, Q^{L} \qquad (4.37)$$

satisfies the spin $\frac{1}{2}(n - 1)$ zero rest-mass equation whenever $\phi_{A\ldots L}$ satisfies the spin $\frac{1}{2}n$ equation (3.1); and furthermore

$$\psi_{A'\ldots K'L'} = - i\, Q^{L}\, \nabla_{LL'}\, \chi_{A'\ldots K'} - n\, \chi_{(A'\ldots K'}Q_{L')} \qquad (4.38)$$

satisfies the spin $\frac{1}{2}n$ equation (3.3) whenever $\chi_{A'\ldots K'}$ satisfies the spin $\frac{1}{2}(n - 1)$ equation, where the constant spinor $Q_{A'}$ is related to Q^{A} by (2.14):

$$\nabla_{AA'}\, Q^{B} = - i\varepsilon_{A}{}^{B}\, Q_{A'} \quad , \qquad (4.39)$$

so that the twistor Q^{α} is given by

$$Q^{\alpha} \longleftrightarrow (Q^{A}, Q_{A'}) \qquad (4.40)$$

If $f(W_{\alpha})$ and $h(Z^{\alpha})$, homogeneous of degrees $- n - 2$ and $- n - 1$ respectively, generate $\phi_{A\ldots L}$ and $\chi_{A\ldots K}$, one may verify directly that

$$(Q^{\alpha}W_{\alpha})f \quad \text{and} \quad Q^{\alpha}\partial h/\partial Z^{\alpha} \qquad (4.41)$$

will generate $\gamma_{A\ldots K}$ and $\psi_{A\ldots M}$, respectively. Note that $Q^{\alpha}W_{\alpha}$ and $Q^{\alpha}\partial/\partial Z^{\alpha}$ are helicity raising operators. Correspondingly, $\bar{Q}_{\alpha}Z^{\alpha}$ and $\bar{Q}_{\alpha}\partial/\partial W_{\alpha}$ are helicity lowering operators.

From (4.37) and (4.38) we shall derive

$$\langle\bar{\psi}|\phi\rangle = \langle\bar{\chi}|\gamma\rangle \tag{4.42}$$

To establish this for (4.34), we first note that a set of potentials for $\chi_{A'\ldots K'}$ satisfying (4.35) can be related to those for $\psi_{A'\ldots L'}$ (cf. (4.38)) by

$$ {}^{(r)}\psi^{G\ldots L}_{A'\ldots E'F'} = -\,iQ^{M}\,\nabla_{MF'}\;{}^{(r-1)}\chi^{G\ldots L}_{A'\ldots E'} - r\;{}^{(r-1)}\chi^{G\ldots L}_{(A'\ldots E'}\,Q_{F')}\;.$$

Substituting this into (4.34) we get

$$\langle\bar{\psi}|\phi\rangle = i^{n-1}\,k\oint\{\,-i\;\phi_{AB\ldots L}\,Q^{M}\nabla_{MA'}\;{}^{(o)}\chi^{B\ldots L} - \phi_{AB\ldots L}\;{}^{(o)}\chi^{B\ldots L}Q_{A'}\}d^{3}x^{AA'}$$

which is equal to

$$\langle\bar{\chi}|\gamma\rangle = i^{n-2}\,k\oint\phi_{AB\ldots KL}\,Q^{L}\;{}^{(1)}\chi^{B\ldots K}_{A'}\;d^{3}x^{AA'}$$

by virtue of

$$Q^{L}\;{}^{(1)}\chi^{B\ldots K}_{A'} = Q^{L}\,\nabla_{MA'}\;{}^{(o)}\chi^{B\ldots KM} = Q^{M}\,\nabla_{MA'}\;{}^{(o)}\chi^{B\ldots KL} - Q^{M}\,\nabla^{L}_{A'}\;{}^{(o)}\chi^{B\ldots K}{}_{M}$$

and

$$\oint \nabla^{L}_{A'} \{\phi_{AB\ldots KL}\, Q^{M}\, \overset{(o)}{\chi}{}^{B\ldots K}{}_{M}\}\, d^3x^{AA'} = 0 \quad ,$$

this last relation being a consequence of the fact that

$$\nabla^{L}_{A'}\, \alpha_{AL}\, d^3x^{AA'} = d\,\{\tfrac{i}{2}\, \alpha_{AB}\, dx^{A}_{A'} \wedge dx^{BA'}\} \qquad (4.43)$$

whenever $\alpha_{AL} = \alpha_{LA}$. The proof for (4.33) is similar.

Using the same bracket notation for the scalar product when it is the twistor functions themselves involved (so (4.30) can be written $\langle\bar\psi|\phi\rangle = \langle\bar g|f\rangle$) we can re-express (4.42) as $\langle Q^\alpha\, \overline{\partial h/\partial Z^\alpha}|f\rangle = \langle\bar h|Q^\alpha\, W_\alpha f\rangle$ (cf. (4.41)) or, because of the linearity of the scalar product, as

$$\langle\overline{\partial h/\partial Z^\alpha}|f\rangle = \langle\bar h|W_\alpha f\rangle \qquad (4.44)$$

or as

$$\langle\bar h|\ \overleftarrow{\frac{\partial}{\partial Z^\alpha}}\ |f\rangle = \langle\bar h|W_\alpha|f\rangle \qquad (4.45)$$

Taking complex conjugates and relabling $\bar W^\alpha$ as Z^α, relation (4.45) expresses the operator equation $\partial/\partial\bar Z_\alpha = Z^\alpha$, or, because(41) the derivative acting to the left in (4.45) must count as the operator $-\,\partial/\partial Z^\alpha$, (4.45) directly expresses the operator equation $-\,\partial/\partial Z^\alpha = \bar Z_\alpha$ (cf. (3.35)).

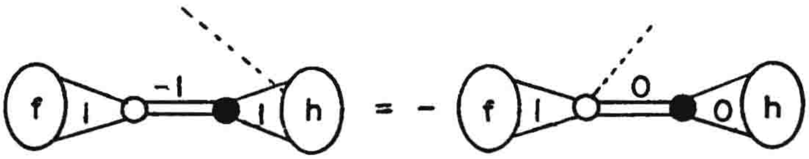
Equation (4.44) can be used to derive the twistor expressions for the scalar product for arbitrary spin massless particles from those of spin $\frac{1}{2}$. (For free waves, there is no loss of generality in the passage from $\phi_{A\ldots L}$ to $\gamma_{A\ldots K}$ in (4.37) or in the passage from $\chi_{A'\ldots K'}$ to $\psi_{A'\ldots L'}$ in (4.38)$^{(28)}$). Take the case of spin zero first. Let $h(Z^\alpha)$ and $f(W_\alpha)$ be of respective homogeneity degrees -2 and -3. Then the scalar product on the right in (4.44) is the expression for spin zero, while that on the left is for spin $\frac{1}{2}$. We have, by (4.44) and (4.30),

$$\langle\bar{h}|W_\alpha f\rangle = ik'\oint \frac{f(W_\sigma)}{W_\beta Z^\beta}\,\frac{\partial h(Z^\rho)}{\partial Z^\alpha}\,d^4W \wedge d^4Z$$

$$= -\,ik'\oint \frac{W_\alpha f(W_\sigma)\,h(Z^\rho)}{-(W_\beta Z^\beta)^2}\,d^4W \wedge d^4Z \qquad (4.46)$$

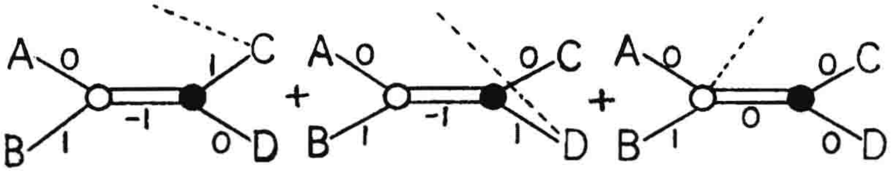
integrating by parts with respect to Z^α. This tells us that the twistor scalar product for spin zero differs from the spin $\frac{1}{2}$ formula (4.30) simply in that the factor $W_\beta Z^\beta$ in the denominator in (4.30) must be replaced by its square.

The diagram for the integration by parts procedure (4.46) can be given as



(4.47)

where the dotted line joined to the h circle refers to the operation $\partial/\partial Z^\alpha$ on h, while the dotted line connected to the white vertex denotes the factor W_α in the numerator. In terms of elementary states, with f and h being given as in (4.13) with $p = 0$, $q = 1$, $r = 0$, $s = 0$, we can re-express (4.47) as



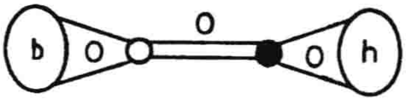
$= 0$ (4.48)

(the dotted line connected to C denoting the presence of a factor C_α, etc.) which states the expansion of

$$\oint \frac{1}{A^{\beta}W_{\beta}(B^{\gamma}W_{\gamma})^{2}} \frac{\partial}{\partial Z^{\alpha}} \{ \frac{1}{W_{\rho}Z^{\rho}C_{\sigma}Z^{\sigma}D_{\tau}Z^{\tau}} \} \mathbb{D}\, WZ = 0.$$

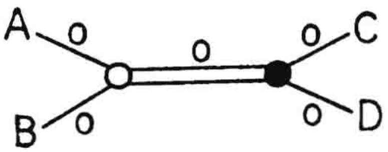
Equation (4.48) is typical of the diagrammatic representation of integration by parts. We shall see more of this later.

Denoting $(Q^{\alpha}W_{\alpha})f$ by b, we see, from (4.47) that the diagram for the scalar product $\langle\bar{h}|b\rangle$ is (ignoring an overall factor)



(4.49)

which, in terms of elementary states ($p = q = r = s = 0$), is



(4.50)

A similar procedure of integration by parts can be used to obtain the twistor expression for higher spin scalar products. But now there is an essential difference owing to the fact that the term $(W_\alpha Z^\alpha)^{-1}$ has to be integrated with respect to Z^α. Methods for achieving this include employing logarithmic terms (but the non-homogeneity of the logarithm function can cause some problems) or using non-integral exponents and taking the appropriate limit. In either case some difficulties may arise owing to the occurrence of many sheeted Riemann surfaces. The method I am presenting here avoids these problems by adopting Sparling's suggestion to use contours with boundary. The idea is to interpret the integration by parts,

$$\oint (W_\beta Z^\beta)_1 \, W_\alpha f \, d^4Z = \oint \frac{W_\alpha f}{2\pi i \, W_\beta Z^\beta} \, d^4Z = - \oint (W_\beta Z^\beta)_0 \, \frac{\partial f}{\partial Z^\alpha} \, d^4Z \qquad (4.51)$$

so that the symbol $(W_\beta Z^\beta)_0$ does not indicate a singularity that is to be integrated around but is a formal symbol which converts the meaning of the symbol $\oint$ into an open contour integral ($\rlap{\,c}\int$) where the boundary of the contour lies in $W_\beta Z^\beta = 0$, the numerical value of the factor $(W_\beta Z^\beta)_0$ being unity. (Note that this is consistent with (4.3) and (4.6).) To establish (4.51), assuming that suitable contours exist, we observe

$$\begin{aligned} \oint (W_\beta Z^\beta)_0 \, \frac{\partial f}{\partial Z^\alpha} \, d^4Z &= \rlap{\,c}\int \frac{\partial f}{\partial Z^\alpha} \, d^4Z \\ &= \rlap{\,c}\int d(f \, d^3Z_\alpha) = \oint_{W_\beta Z^\beta = 0} f \, d^3Z_\alpha \,, \end{aligned} \qquad (4.52)$$

by the divergence theorem, where

$$d^3Z_\alpha = \frac{1}{6}\,\varepsilon_{\alpha\beta\gamma\delta}\;dZ^\beta \wedge dZ^\gamma \wedge dZ^\delta \tag{4.53}$$

The final integral in (4.52), being taken over the complex hypersurface $W_\beta Z^\beta = 0$ in Z-space, can be converted back to a 4-dimensional integral by employing the same device that was used in (4.23), namely to use Cauchy's theorem to express what is, in effect, a δ-function in one of the Z-variables. We have

$$d^3Z_\alpha \wedge d(W_\beta Z^\beta) = -\,W_\alpha\, d^4Z \quad (\text{on } W_\beta = \text{const.}) \tag{4.54}$$

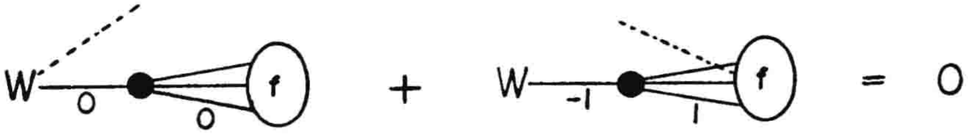
so we can express (4.52) as

$$\oint\limits_{W_\beta Z^\beta=0} f\, d^3Z_\alpha = \oint f\, d^3Z_\alpha \wedge \left\{\frac{d(W_\beta Z^\beta)}{2\pi i\, W_\gamma Z^\gamma}\right\}$$

$$= -\oint \frac{W_\alpha f}{2\pi i\, W_\gamma Z^\gamma}\quad d^4Z \tag{4.55}$$

which, when combined with (4.52), yields (4.51) as required.

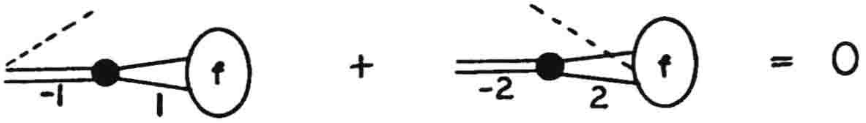
The diagrammatic form of (4.51) is



(4.56)

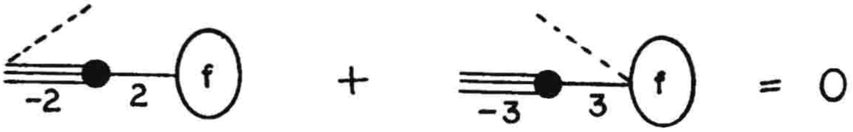
although the assignment of multiplicities to the edges that is given in

(4.56) is somewhat arbitrary. We could as well express the relationship in (4.56) as



(4.57)

or



(4.58)

The multiplicities would be intended to convey some information concerning the nature of the contour, but as yet nothing has been said about how the contours in (4.52) and (4.55) are to be chosen (except that the bounded contour has to have boundary in $W_\alpha Z^\alpha = 0$). Indeed, (4.51) is really too special as it stands, since the twistor W_α is taken as a constant throughout the integration. The application of this result to twistor diagrams frequently requires contours in which Z^α cannot always be varied independently of other variables such as W_α. So (4.51) should be more correctly re-expressed as

$$\oint (W_\beta Z^\beta)_1 \, W_\alpha f(Z^\gamma, W_\gamma, \dots) \, \mathbb{D} ZW\dots = -\oint (W_\beta Z^\beta)_0 \, \frac{\partial}{\partial Z^\alpha} f(Z^\alpha, \dots) \, \mathbb{D} ZW\dots \quad . \tag{4.59}$$

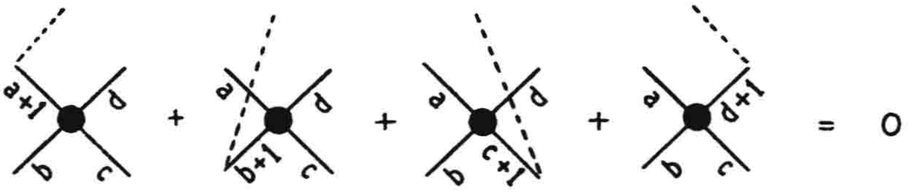
In fact contours are likely not to exist for which the strict form

(4.51) is valid. But we can regard the expression as a shorthand for (4.59), for which suitable contours can be found when the singularities of f are appropriate. The question of when suitable contours exist is a purely topological one and the reader is referred to the general discussion in Sparling's article.(38)

We can extend (4.59) by repeated integration by parts to

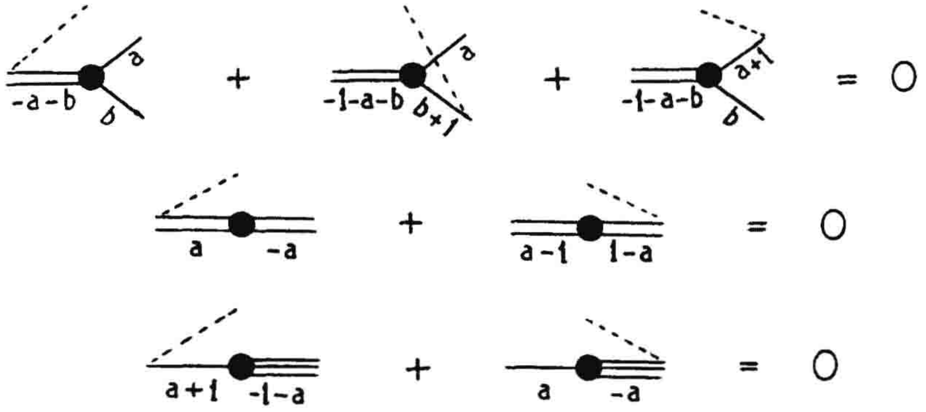
$$\oint (W_\beta Z^\beta)_{1-p}\, W_\alpha\, f\, \mathbb{D}ZW\ldots = -\oint (W_\beta Z^\beta)_{-p}\, \frac{\partial f}{\partial Z^\alpha}\, \mathbb{D}ZW\ldots \qquad (4.60)$$

where the definition of $(W_\beta Z^\beta)_{-p}$ for p = 1,2,3,... differs from that of $(W_\beta Z^\beta)_0$ only in that an additional factor $(W_\beta Z^\beta)^p/p!$ is included (cf. (4.6), (4.3)). If p = 0, -1, -2, ... (or is non-integral) then the factor $(W_\beta Z^\beta)_{-p}$ is given by (4.4), having a singularity which is to be integrated around a closed contour in the usual way. Then (4.60) remains valid (whenever the topology allows suitable contours to exist). When applied in a twistor diagram, (4.60) yields (cf. (4.48))



(4.61)

where a + b + c + d = - 1, and in more 'degenerate' cases

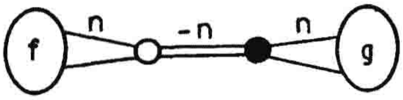


(4.63)

An extra term could also appear if there were a given dotted line attached to the black vertex. Applying (4.60) repeatedly to the twistor scalar product formulae (4.30), (4.46) we obtain, for spin $\frac{1}{2}n$,

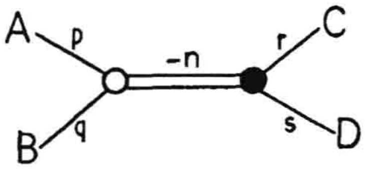
$$\langle \bar{g} | f \rangle = (i)^{2n-1} k'' \oint f(W_\alpha)\, (W_\beta Z^\beta)_{2-n}\, g(Z^\gamma)\, \mathbb{D}WZ \tag{4.64}$$

with diagram



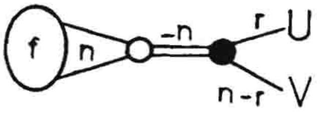
(4.65)

or, in the case of elementary states, with p + q = n = r + s,



(4.66)

It may be mentioned, in passing, that there is an alternative formula to (3.20) for obtaining the components of the massless field $\phi_{A\ldots L}$ from its twistor function f, namely the scalar product expression intermediate between (4.65) and (4.66):

$\phi_r(U_\alpha\ ,V_\alpha)$ =  $(r = 0,\ldots,n)$,

which is useful in some contexts. This may be derived by expressing f as a linear combination of elementary states and explicitly evaluating the integral (4.66) or, following an argument due to J.R. Moore, by repeatedly integrating by parts and using a result derived from (4.72) to obtain the result for $n = r = 0$.

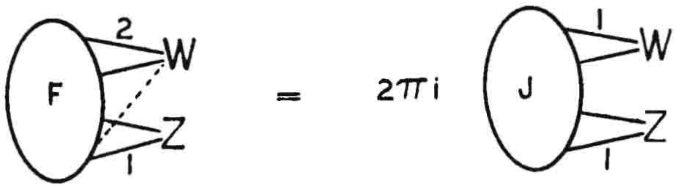
It is worth pointing out that the integration by parts procedure can also be used to relate the two twistor expressions (4.17) and (4.26) that have been obtained for the charge of a system. (The following discussion is based, to a large extent, on an argument due to J.R. Moore.) Maxwell's equations, when sources are present, have the spinor form

$$\nabla^B_{A'}\ \phi_{AB} = -\ 2\pi j_{AA'} \tag{4.67}$$

Since ϕ_{AB} does not now satisfy (3.1) we need a two variable twistor function $F(W_\alpha, Z^\alpha)$, homogeneous of degrees $-\ 4,\ -\ 2,$ in order to generate the field. Comparing (4.67) with (3.43) we see that a twistor function $J(W_\alpha, Z^\alpha)$ which generates j_a can be obtained using (3.45):

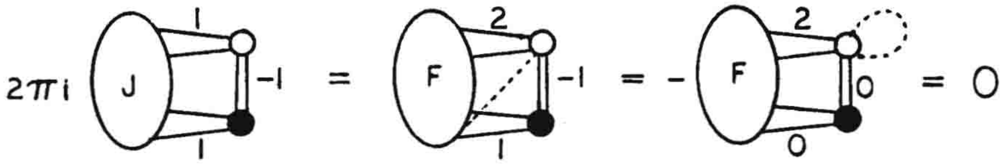
$$I^{\alpha\beta}\ W_\alpha \frac{\partial F}{\partial Z^\beta} = 2\pi i\ J \tag{4.68}$$

In terms of diagrams this is



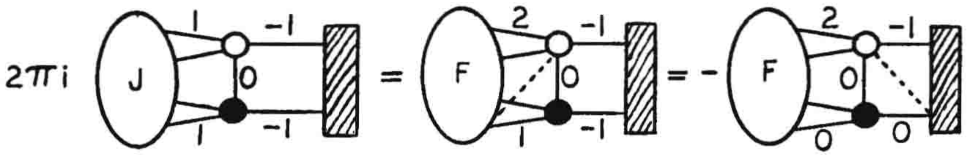
(4.69)

(the dotted line on the left in effect joins two 'white' vertices and so an infinity twistor is involved (cf. fig. 6.7)). If the charge integral (4.26) had been over a compact contour, then its diagram would be proportional to



(4.70)

since the dotted line joined in a loop involves the factor $I^{\alpha\beta}W_\alpha W_\beta = 0$ ($I^{\alpha\beta}$ being skew). This illustrates the fact that the total charge for a closed universe (in this case compactified Minkowski space) must vanish. For a non-zero charge we require a bounded contour (corresponding to the fact that a neighbourhood of the point I is being omitted from the spacelike hypersurface *S* of integration). We can express this diagrammatically



(4.71)

where the shaded rectangle denotes the (holomorphic) locus L in

(W_α, Z^β)-space where the boundary of the contour is to reside. A convenient choice for L is as defined by $W_\alpha L^\alpha_\beta Z^\beta = 0$ for some twistor L^α_β. This includes the case when ∂S is a spacelike 2-sphere. Making this choice and assuming that the integral on the right of (4.71) can be performed by carrying out the Z-integral for each constant W_α first and then performing the W-integral, we can express this Z-integral as

$$\oint \left\{ \frac{W_\alpha U_\beta I^{\alpha\beta}}{W_\rho Z^\rho U_\sigma Z^\sigma} \right\} F(W_\gamma, Z^\gamma) \quad d^4 Z \qquad (4.72)$$

where U_σ is here given by $U_\sigma = L^\rho_\sigma W_\rho$.

The expression (4.72) has another significance. Suppose that F is not assumed to have homogeneity degree -4 in W_γ (although the degree -2 in Z^γ is still relevant), so that it refers also to spins other than 1, and take U_α and W_α to be arbitrary. Let the point y^a of $\mathbb{M}^I$ be the intersection of W_α with U_α (i.e., defined by the twistor $W_{[\alpha} U_{\beta]}$). Then with $Z^\alpha \leftrightarrow (iy^{AA'} Z_{A'}, Z_{A'})$ the integral

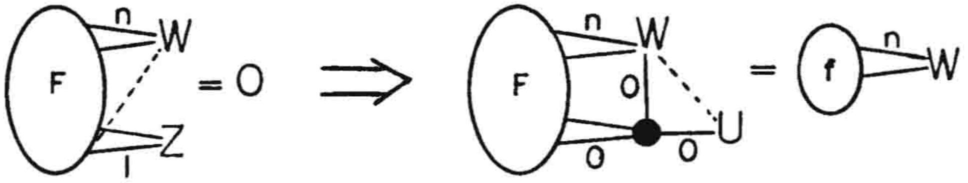
$$\oint F(W_\gamma, Z^\alpha)\, d^2 Z$$

for fixed y^a (and W_γ, U_γ) can, as is not hard to show, be re-expressed as (4.72).

The space-time field generated by F, of homogeneity degrees $(-n-2, -2)$, has n unprimed indices only and satisfies the *spin* $\frac{1}{2}n$ *free-field equations* (3.1) *if*

$$I^{\alpha\beta}W_\alpha \, \partial F/\partial Z^\beta = 0 \tag{4.73}$$

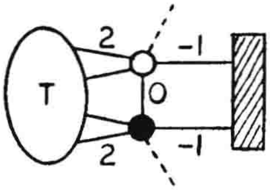
(cf. (3.45)). This condition also expresses the fact that (4.72) is independent of U_σ (where U_σ and W_ρ are now taken to be independent of one another), so (4.72) may be regarded as giving a method of passing from a 2-variable twistor function F for a zero rest-mass free-field to a one-variable twistor function $f(W_\alpha)$ generating the same field. In terms of diagrams:



(4.74)

Even if the free-field equations do not hold, this gives us an $f(W_\alpha)$ provided U_α is given (up to proportionality) as a function of W_α. This indeed was the case in the construction leading to (4.72), and the resulting field which f generates it may be viewed as giving a description of a free field defined at ∂S , which agrees with ϕ_{AB} at ∂S . Assuming that the (basically topological) questions concerning contour location can be sorted out (including the fact that the final bounded contour for the W_α-integration can be made closed) - and formally, in any case, - the formulae (4.71) and (4.74) do give the required relation between the two twistor expressions (4.17) and (4.26).

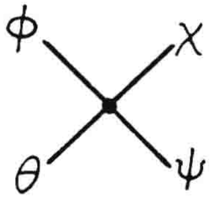
It is possible also to give a corresponding argument for spin 2, namely linear gravitation, where the source T_{ab} is taken to be trace-free (e.g. derived from zero rest-mass fields), this being needed to make $T_{AA'BB'}$ symmetric in A,B and in A',B' so that, since $\nabla^a T_{ab} = 0$, a two-variable twistor function T generates it. The trace-free quantity



$$\text{i.e., } E^{\alpha}_{\beta} = g \oint \frac{W_\beta Z^\alpha}{W_\gamma Z^\gamma} T(W_\sigma, Z^\sigma)\, \mathbb{D} WZ \qquad (4.75)$$

(with g some constant) defines the 15 conserved quantities arising from the conformal generators (cf. (3.59), (3.60)), the 10 components of energy-momentum and angular momentum being singled out by (3.57). Indeed, the analogous calculation to the electromagnetic one described above does, in fact, yield the required association (3.57) between the integrals (4.18) and (4.75).

Finally let us briefly consider some interactions. In the absence of a coherent method of generating interactions generally, motivated from within twistor theory itself, one procedure has been simply to translate the position space Feynman diagrams, corresponding to some perturbation expansion, into twistor terms. Consider first a single Feynman vertex for a massless scalar ϕ^4 interaction:



(4.76)

We have to compute the space-time integral

$$\oint \phi(x^a)\ \theta(x^a)\ \psi(x^a)\ \chi(x^a)\ \ d^4x \tag{4.77}$$

taken over the (compact) space $\mathbb{M}$, or else over some suitable 4-dimensional closed contour in $\mathbb{CM}$, where

$$d^4x = dx^0 \wedge dx^1 \wedge dx^2 \wedge dx^3 = \frac{1}{24}\, e_{abcf}\ dx^a \wedge dx^b \wedge dx^c \wedge dx^f. \tag{4.78}$$

The translation turns out simplest if we take ϕ, θ, ψ and χ to be generated by twistor functions $f(W_\alpha)$, $t(X^\alpha)$, $p(Y_\alpha)$ and $h(Z^\alpha)$, two of which are defined on $\begin{bmatrix}1\\0\end{bmatrix}$ -twistor space and two on $\begin{bmatrix}0\\1\end{bmatrix}$ -space, all the functions being homogeneous of degree − 2 (cf. (3.14)). The twistor expression for (4.77) is then

$$\oint f(W_\alpha)\ t(X^\alpha)\ p(Y_\alpha)\ h(Z^\alpha)\ d^2W \wedge d^2X \wedge d^2Y \wedge d^2Z \wedge d^4x \tag{4.79}$$

but where each of $W_\alpha,\ldots,Z^\alpha$ is constrained to pass through the same point x^a. However, for a proper twistor integral, each of $W_\alpha,\ldots,Z^\alpha$ has to be allowed to vary independently. The constraint, here, that they all pass through the same point can be expressed as

$$W_\alpha X^\alpha = Y_\alpha X^\alpha = Y_\alpha Z^\alpha = W_\alpha Z^\alpha = 0 \tag{4.80}$$

(If one thinks in terms of incidence between real null lines this condition appears inadequate, but this is to ignore the fact that x^a can be complex. It is easier to think in terms of points and planes in $\mathbb{PT}$; (4.8 states that the join of points $\underset{\sim}{X}$ and $\underset{\sim}{Z}$ in $\mathbb{PT}$ coincides with the intersection of planes $\underset{\sim}{W}$ and $\underset{\sim}{Y}$ in $\mathbb{PT}$ - assuming $\underset{\sim}{X} \neq \underset{\sim}{Z}$ and $\underset{\sim}{W} \neq \underset{\sim}{Y}$.) Thus, as in (4.23) and (4.55), we must incorporate the reciprocal of each scalar product (4.80), together with the differential of each scalar product, into the integral (4.79). For this, we require the identity

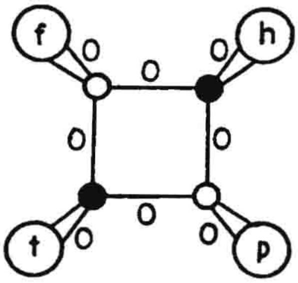
$$d^2W \wedge d^2X \wedge d^2Y \wedge d^2Z \wedge d^4x \wedge d(W_\alpha X^\alpha) \wedge d(Y_\beta X^\beta) \wedge d(Y_\rho Z^\rho) \wedge d(W_\sigma Z^\sigma)$$

$$= i\ d^4W \wedge d^4X \wedge d^4Y \wedge d^4Z \qquad (4.81)$$

where x^a is defined as the intersection of W_α and Y_α, and y^a as the intersection of X^α and Z^α, i.e.,

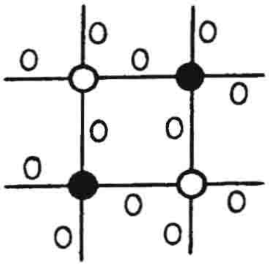
$$\begin{aligned} W_\alpha &\longleftrightarrow (W_A, -ix^{AA'} W_A), & Y_\alpha &\longleftrightarrow (Y_A, -ix^{AA'} Y_A) \ , \\ X^\alpha &\longleftrightarrow (iy^{AA'} X_{A'}, X_{A'}) \ , & Z^\alpha &\longleftrightarrow (iy^{AA'} Z_{A'}, Z_{A'}) \ . \end{aligned} \qquad (4.82)$$

The conclusion, expressed diagrammatically, is that (4.77) is proportional to



(4.83)

or, in terms of elementary states, to



(4.84)

which we may regard as the twistor version* of the Feynman ϕ^4 vertex (4.76).

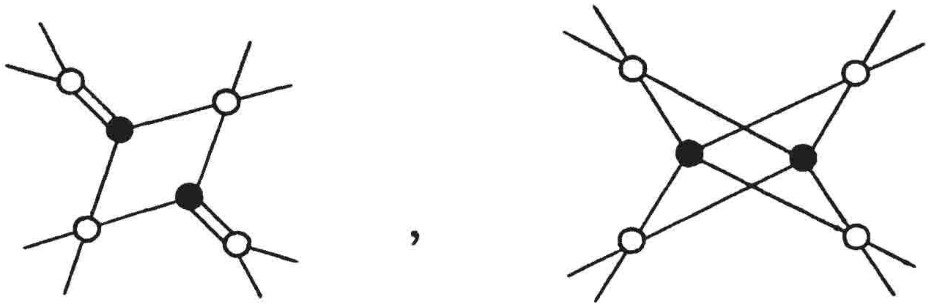
It turns out, in fact, that there is more than one contour of interest for (4.84) (see Sparling's article). There is one which gives an answer symmetrical in all four elementary states and (apart from the misprint** in the sign of the exponent and the omission of an overall factor) is as given in ref. (19). However, one may say that this answer is not the 'physical' one since it becomes infinite for certain special relative locations between the elementary states (because the contour 'pinches'). Instead, one may choose a contour which, for a given set of elementary states, corresponds to taking (4.77) over the real points of $\mathbb{M}$. This gives a finite answer (but it is not symmetrical in all four elementary states). It is evident from this that finding the integrand for a twistor diagram is only one part of the problem. One needs a method also of defining the topological location (homology class) of the contour.

---

* Note that we are performing a 16-dimensional integral (with a closed loop) to compute a process which in the usual Feynman momentum space representation would not require integration at all! It is important, in twistor theory, not to be daunted by such matters!

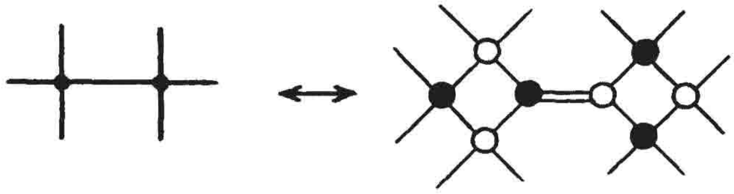
** It is worth pointing out that there is a misprint also in equation (4.26) of ref. (19). The right-hand side should read $\alpha^{a}{}_{\beta}{}^{c}{}_{\lambda}{}^{b}{}_{\mu}{}^{d}{}_{\delta}$.

There are also other more complicated ways of expressing (4.77) in twistor terms. For example, if all the states are described by twistor functions defined on the $\begin{bmatrix}0\\1\end{bmatrix}$-twistor space, then alternative forms such as



(4.85)

may be derived (where all lines are now understood to bear the label '0'). Note that the presence of closed loops in twistor diagrams does not give rise to infinities - naturally enough, in the above cases, since merely a single Feynman vertex is being depicted. But this seems to be true also for higher order diagrams, so the relation between (divergent) Feynman diagrams and twistor diagrams is not completely clear. Nevertheless it seems to be legitimate to string twistor diagrams together in the obvious way, e.g.,

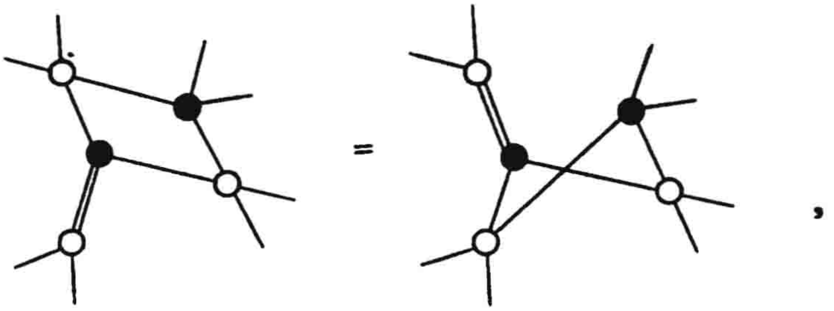


(Feynman) (twistor)

(4.86)

and, as far as one can attribute a meaning to it, the central double line in the twistor diagram represents a virtual particle 'off the mass shell'.

There are many identities relating twistor diagrams, such as (again with zero on all lines)



(4.87)

but which do not hold for all choices of contour. This emphasizes once again, the importance of the contour topology as an integral part of the theory.

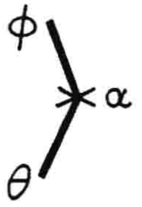
To end the section, I shall consider Möller scattering (for massless electrons). The diagram will be built up in steps.* Let ϕ_A and $\theta_{A'}$ be massless spin $\frac{1}{2}$ fields, generated by twistor functions $f(W_\alpha)$ and $t(X^\alpha)$, each homogeneous of degree -3, and let α_a be an electromagnetic potential generated by a twistor function $A(Y_\alpha, Z^\alpha)$ homogeneous of degrees -3, -3. We wish first to translate the integral

$$\oint \phi(x^a)_B \ \theta(x^a)_{C'} \ \alpha(x^a)^{BC'} \ d^4x \tag{4.88}$$

(taken over $\mathbb{M}$), corresponding to the Feynman graph

---

* Part of this argument is due to J.R. Moore.



, (4.89)

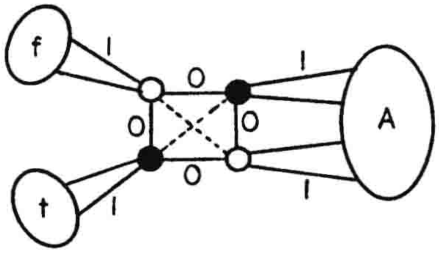
into a twistor form. Substituting the twistor functions into (4.88) we get

$$\oint W_B\, f(W_\beta)\, X_{C'}\, t(X^\gamma)\, Y^B Z^{C'}\, A(Y_\rho, Z^\rho)\, d^2W \wedge d^2X \wedge d^2Y \wedge d^2Z \wedge d^4x.$$

Noting that

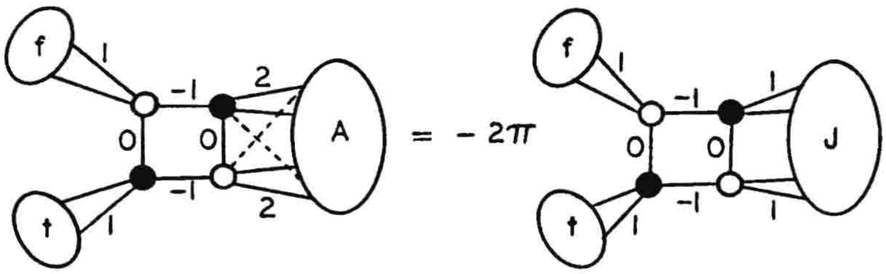
$$W_B Y^B = W_\alpha Y_\beta I^{\alpha\beta} \ \underline{\text{and}}\ X_{C'} Z^{C'} = X^\alpha Z^\beta I_{\alpha\beta}$$

(cf. (2.44)) we can follow the same argument leading from (4.79), to obtain the twistor diagram



(4.90)

to represent (4.89). Integrating by parts twice (and noting that no term arises from the dotted lines joined to one another owing to the presence of infinity twistors, and (2.45)), we derive



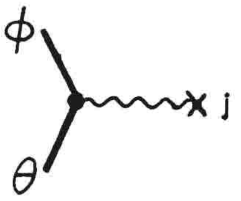
(4.91)

where the potential is now related to a current source by

$\Box\,\alpha_a = 4\pi\, j_a$, generated by

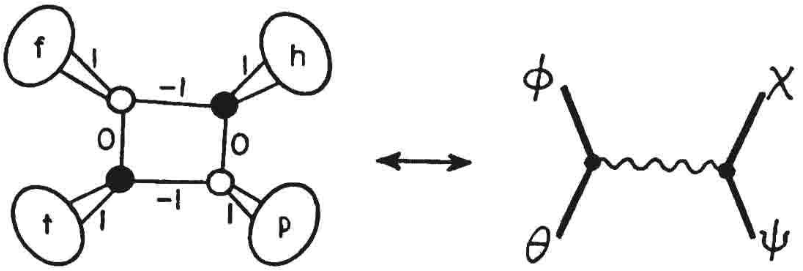
$$- 2I_{\alpha\beta}\, Z^{\alpha}\, \frac{\partial}{\partial Y_{\beta}}\, \{I^{\rho\sigma}\, Y_{\rho}\, \frac{\partial A}{\partial Z^{\sigma}}\} = 4\pi J \qquad (4.92)$$

(cf. (3.47)). Owing to the inversion of the D'Alembertian that is involved in passing from j_a to α_a, the diagram on the right of (4.91) may be regarded as representing the Feynman diagram



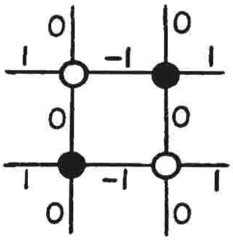
(4.93)

Finally, substituting $j_{AA'} = \psi_A \chi_{A'}$ where ψ_A and $\chi_{A'}$ are massless spin $\frac{1}{2}$ fields, generated by respective twistor functions $p(Y_\alpha)$ and $h(Z^\alpha)$ each of homogeneity degree -3, we can re-express (4.91) and (4.93) as



(4.94)

to give the twistor version(19) of Möller scattering between massless electrons. In terms of elementary states this is



(4.95)

5. CURVED-SPACE TWISTORS

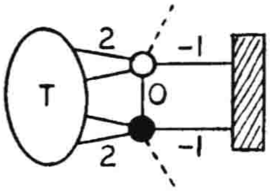
There are several different, but inter-related ways of adapting twistor theory to curved space-time. I shall describe four different notions of twistor, and show how they relate to one another. These are

(a) local twistors,
(b) global null twistors,
(c) hypersurface twistors,
(d) asymptotic twistors.

It would clearly have been more satisfactory had there been just one 'manifestly correct' way of handling twistors in curved space. The difficulty is that most of the ways of describing flat-space twistors do not work when the space becomes curved. For example; there is no satisfactory definition of the angular momentum of a system about a space-time point in general relativity, so a definition like (2.11) seems unsuitable. One might think of adopting the twistor equation (2.13), as an alternative approach, where $\nabla_{BB'}$ denotes covariant derivative. But there arises a consistency condition $0 = \nabla^{B'(C}\nabla^{B}_{B'}\omega^{A)} = -\Psi^{ABC}{}_{D}\,\omega^{D}$, where Ψ_{ABCD} is the conformal spinor, related to Weyl's conformal tensor C_{abcd} according to (3.8). This implies that either $\omega^A = 0$, or else that the conformal tensor is of a very special type (referred to as 'null') and, in addition, ω^A is related to it in a very special way (with 'flagpole' in the direction of the fourfold principal null direction of Ψ_{ABCD}). Even if all this is satisfied, (2.13) will admit a non-zero solution only in exceptional cases - and when it does, unless $\Psi_{ABCD} = 0$, this solution will be unique up to proportionality. Thus only when $\Psi_{ABCD} = 0$, which is the condition that the space-time be conformally flat, can the twistor equation (2.13) provide a satisfactory definition of a general twistor. Furthermore there appears to be no modification of (2.13), having solutions defined throughout space-time, which could provide such a definition.

Another possibility might be to consider null twistors only and to make the obvious generalization from a null straight line Z in $\mathbb{M}^I$ to a null geodesic Z in the curved space-time M. The $\pi_{A'}$-spinor, which determines the phase and momentum scaling of the twistor, must then be taken to be defined along Z with the momentum vector $p_{AA'} = \bar{\pi}_A \pi_{A'}$ tangent to Z, the spinor $\pi_{A'}$ being parallelly propagated along Z. (This is conformally invariant, cf. (2.17).) Indeed, this is actually the definition of a global null twistor in M. The difficulties with this, however, are that there seems no unique way of extending the definition to non-null twistors and (related to this) no unique way of assigning a complex structure to the global twistor space even if the non-null twistors have in some way been given. The geometric reasons for these difficulties, and their partial resolution in terms of the concepts of hypersurface twistors and asymptotic twistors, will be given in a little detail shortly. The problem is one which also confronts another approach that might be considered for introducing twistors into curved space-time, namely to look for analogues of the complex totally null

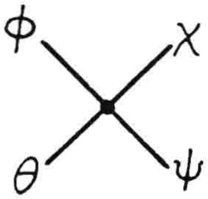
It is possible also to give a corresponding argument for spin 2, namely linear gravitation, where the source T_{ab} is taken to be trace-free (e.g. derived from zero rest-mass fields), this being needed to make $T_{AA'BB'}$ symmetric in A,B and in A',B' so that, since $\nabla^a T_{ab} = 0$, a two-variable twistor function T generates it. The trace-free quantity



$$\text{i.e., } E^{\alpha}_{\beta} = g \oint \frac{W_\beta Z^\alpha}{W_\gamma Z^\gamma} T(W_\sigma, Z^\sigma)\, \mathbb{D}\, WZ \qquad (4.75)$$

(with g some constant) defines the 15 conserved quantities arising from the conformal generators (cf. (3.59), (3.60)), the 10 components of energy-momentum and angular momentum being singled out by (3.57). Indeed, the analogous calculation to the electromagnetic one described above does, in fact, yield the required association (3.57) between the integrals (4.18) and (4.75).

Finally let us briefly consider some interactions. In the absence of a coherent method of generating interactions generally, motivated from within twistor theory itself, one procedure has been simply to translate the position space Feynman diagrams, corresponding to some perturbation expansion, into twistor terms. Consider first a single Feynman vertex for a massless scalar ϕ^4 interaction:



(4.76)

Local twistors are not to be thought of as providing the sought-for generalization of flat-space twistors to M. Each local twistor is defined at a point of M, so we cannot use local twistors as something more primitive from which the points of M are to be themselves constructed. Nonetheless, local twistors provide a well-defined conformally invariant calculus which is useful, particularly, for the discussion of asymptotic twistors.

A local twistor Z^α at a point $P \in M$ can be described by a spinor pair ω^A, $\pi_{A'}$ at P

$$Z^\alpha \longleftrightarrow (\omega^A, \pi_{A'}) \tag{5.1}$$

with respect to the metric g_{ab} on M. If the metric is conformally rescaled according to (2.15), (2.16) then the representation of Z^α must be changed according to (2.19):

$$Z^\alpha \longleftrightarrow (\hat{\omega}^A, \hat{\pi}_{A'}) = (\omega^A, \pi_{A'} + iT_{AA'}\,\omega^A) \tag{5.2}$$

with T_a given by (2.18). In order to compare local twistors at different points of M, we require the concept of local twistor transport along a curve τ in M with tangent vector t^a

$$\begin{aligned} t^b\,\nabla_b\,\omega^A &= -\,it^{AB'}\ \pi_{B'} \\ t^b\,\nabla_b\,\pi_{A'} &= -\,i\,P_{BB'AA'}\,t^{BB'}\,\omega^A\,, \end{aligned} \tag{5.3}$$

where

$$P_{ab} = \frac{1}{12} R\, g_{ab} - \frac{1}{2} R_{ab} \quad ,$$

the Ricci tensor $R_{ab} = R_{acb}{}^{c}$ having sign defined according to $R_{abc}{}^{d}\, v^{c} = (\nabla_a \nabla_b - \nabla_b \nabla_a)\, v^{d}$. It should be made clear that the operation of local twistor transport along τ is not really to be thought of as moving the twistor itself along τ, but rather as moving the point ('origin') with respect to which the twistor is referred.

More generally, we can define the covariant derivative, in the direction t^a of a local twistor field on *M* according to

$$t^b \nabla_b Z^\alpha \longleftrightarrow (t^b \nabla_b \omega^A + i t^{AB'} \pi_{B'},\ t^b \nabla_b \pi_{A'} + i\, P_{BB'AA'} t^{BB'} \omega^A). \quad (5.4)$$

It may be verified that this concept of covariant derivative is conformally invariant: if the metric is conformally rescaled according to (2.15) and the description of Z^α undergoes (5.2), then the covariant derivative of Z^α also undergoes (5.2). Thus, in particular, local twistor transport is also conformally invariant.

If we compare (5.4) with (2.14) we see that in flat space-time $\mathbb{M}^I$, the local twistors which are constant correspond precisely to the original twistors of §2. Local twistor transport then defines how the representation of a fixed twistor in terms of a pair of spinors varies as the point of reference (origin) is moved. From the conformal invariance it follows that this must also hold if M is only conformally flat. If M is conformally curved, however, one cannot expect to find many constant non-zero local twistors. If a local twistor (description) is carried around a small loop by local twistor transport, then it will not, in general, return to its original state. In fact one can derive

$$[t^p\nabla_p, u^q\nabla_q]\, Z^\beta - [t,u]^p\nabla_p Z^\beta$$

$$\leftrightarrow t^p u^q \{\varepsilon_{P'Q'}\Psi_{PQA}{}^{B}\omega^A, -i(\varepsilon_{PQ}\nabla_{\mathbf{AA'}}\bar{\Psi}_{B'P'Q'}{}^{A'} + \varepsilon_{P'Q'}\nabla_{\mathbf{BB'}}\Psi_{APQ}{}^{B})\omega^A$$

$$- \varepsilon_{PQ}\Psi_{P'Q'B'}{}^{A'}\pi_{A'}\} \qquad (5.5)$$

where $[t,u]^b$ denotes the Lie bracket $(t^c\nabla_c u^b - u^c\nabla_c t^b)$, showing, once again, that the integrability condition for twistors to be defined globally on M comes out as $\Psi_{ABCD} = 0$, the condition for conformal flatness.

We can also apply the concept of local twistor transport or covariant derivative to a $\begin{bmatrix}0\\1\end{bmatrix}$-twistor W_α, simply by taking the complex conjugates of (5.3) and (5.4). We may verify

$$t^b\nabla_b(Z^\alpha W_\alpha) = Z^\alpha t^b\nabla_b W_\alpha + W_\alpha t^b\nabla_b Z^\alpha \qquad (5.6)$$

the term on the left denoting the ordinary derivative of the scalar

$Z^\alpha W_\alpha$ along τ. From (5.6) we derive the fact that the scalar product between two twistors, for which local twistor transport is maintained along τ, is constant along τ.

Let us now return to the concept of a global null twistor as it was defined above. For the construction to give a reasonable space N for the set of null twistors in M - giving N as a Hausdorff 7-manifold - we require M to be subject to suitable global causality restrictions. For simplicity I shall assume that M is globally hyperbolic(42) and this will suffice. Note that the space $\mathbb{P}N$ of unscaled* null geodesics in M is 5-dimensional; the remaining freedom for $\pi_{A'}$ is just a complex multiplying factor, giving N itself as 7-dimensional.

Now N possesses a closed 2-form $\underline{\omega}$:

$$\underline{\omega} = dp_a \wedge dx^a \tag{5.7}$$

where $p_{AA'} = \bar{\pi}_A \pi_{A'}$. The form (5.7) can be defined initially on the cotangent bundle(43) of M, but it is constant along geodesics (Lagrange identity(43),(5)) and it defines a 2-form on the space of scaled geodesics. Taking this 2-form to be constant under $\pi_{A'} \longmapsto e^{i\theta}\pi_{A'}$, ($\theta$ real), we have $\underline{\omega}$ defined on the space N . The geometrical meaning of $\underline{\omega}$ is that it defines the degree of rotation of a congruence. This can be seen if we rewrite (5.7), for a congruence of geodesics (so p_a becomes a field on M) in the form

---

* An unscaled geodesic may be taken as simply a point set on M. A scaled geodesic carries, in addition, a parallelly propagated tangent vector, i.e., an affine parameter with the origin freedom factored out.

$$\underline{\omega} = \nabla_{[a}\, p_{b]}\, dx^a \wedge dx^b \quad , \tag{5.8}$$

the quantity $\nabla_{[a}\, p_{b]}$ defining the rotation of p_a.

The form (5.7) is the exterior derivative of the 1-form

$$\underline{\phi} = p_a\, dx^a \quad , \tag{5.9}$$

that is,

$$\underline{\omega} = d\underline{\phi} \quad ; \tag{5.10}$$

and $\underline{\phi}$ is also well-defined on the space of scaled null geodesics$^{(5)}$. Again, $\underline{\phi}$ is defined on N and constant under $\pi_{A'} \mapsto e^{i\theta}\pi_{A'}$. The geometrical meaning of $\underline{\phi}$ is that it measures 'time-delay' in a (1-parameter) family of scaled null geodesics. We have $\underline{\phi} = 0$ if the family belongs to a null hypersurface.

In flat space-time we can re-express $\underline{\omega}$ and $\underline{\phi}$ in twistor terms as

$$\begin{aligned} \underline{\omega} &= i\, dZ^\alpha \wedge d\bar{Z}_\alpha \\ \underline{\phi} &= i\, Z^\alpha\, d\bar{Z}_\alpha \end{aligned} \tag{5.11}$$

and this allows us to extend their definition to non-null twistors. An alternative extension for $\underline{\phi}$ would be achieved if we had used $-i\,\bar{Z}_\alpha\, dZ^\alpha$ in place of the above, or else the more symmetrical $\frac{1}{2}(i\,Z^\alpha\, d\bar{Z}_\alpha - i\,\bar{Z}_\alpha\, dZ^\alpha)$, since these each differ from $\underline{\phi}$ by a multiple of $d(Z^\alpha \bar{Z}_\alpha)$, which vanishes on $\mathbb{N}$. The first of these alternatives is, of course, the complex conjugate of $\underline{\phi}$ and the second is its real part. The forms $\underline{\phi}$, $\underline{\omega}$ and the scalar $2s = Z^\alpha \bar{Z}_\alpha$ define what has been referred to⁽³²⁾ as the *invariant structure* of twistor space $\mathbb{T}$. Of course $\underline{\omega}$ is determined by $\underline{\phi}$ according to (5.10), but it has interest in its own right since it defines a *symplectic structure*⁽⁴³⁾ for $\mathbb{T}$. When restricted to the subspace $\mathbb{N}$, $\underline{\omega}$ only defines what is known as a *pre-symplectic* structure, since the matrix it defines is degenerate ($\mathbb{N}$ being odd-dimensional). If we factor out from $\mathbb{N}$ the phase circles

$$Z^\alpha \mapsto e^{i\theta} Z^\alpha \quad (\theta \text{ real}) \tag{5.12}$$

then the space becomes the 6-dimensional space of scaled null geodesics, and $\underline{\omega}$ again defines a symplectic structure on this space.

The reason for the term 'invariant structure' is that, restricted to N at least, this structure has an invariant meaning (as we have seen above) also in the case of *curved* space-times (with N taking over the role of $\mathbb{N}$). We might, for example, have two regions of flat space separated by a region of curvature. The forms $\underline{\omega}$ and $\underline{\phi}$ can be defined on the twistor space in each flat region by means of (5.11) and, because we can re-express them (on N) according to (5.7) and (5.9) and use the known propagation properties of these forms, the definitions in the two flat regions must agree on those parts of N for which both definitions apply. The problem with the non-null parts of the space $\mathbb{T}$ is that, in the general case, no method is known of carrying a non-null twistor across a region of curved space-time from one flat region to another. (Some explicit cases of waves satisfying Einstein's equations are considered in ref. (32) where it does seem to be possible, however.)

In general, it seems that the invariant structure of N, as defined above, is the strongest structure, locally defined in M, that can be given for N which does not have the property of being destroyed by space-time curvature. But a symplectic structure falls a very long way short of the complex structure$^{(43)}$ that is required in order that contour integrals such as those of §3 and §4 can be defined. Without a complex structure, very little of the twistor formalism could survive. It is worthwhile, therefore to examine the space-time meaning of the complex structure in the case of flat twistor space $\mathbb{T}$ and to appreciate the difficulty involved in extending this notion to the curved case.

A complex structure on a manifold may be regarded as a rule for deciding when a complex function defined on an open subset of the manifold is to be called 'holomorphic'. (This rule has to satisfy various properties, such as that a holomorphic function – in the ordinary sense – of any collection of such 'holomorphic' functions is again 'holomorphic'. Locally, one may introduce holomorphic coordinates; then the rule says that the 'holomorphic' functions on open subsets of the manifold are expressible holomorphically in terms of these coordinates.) Thus, we have interpreted the complex structure of $\mathbb{T}$ geometrically as soon as we have a geometric means of recognizing a holomorphic function on $\mathbb{T}$. I shall consider only homogeneous holomorphic functions on $\mathbb{T}$ and give an interpretation of the set of zeros of such a function $f(Z^\alpha)$, restricted to the space $\mathbb{N}$ of null twistors. The null twistors for which $f(Z^\alpha) = 0$ define a congruence of null straight lines in $\mathbb{M}$. The holomorphic nature of f implies, according to the theorem of Kerr(20), that this congruence is shear-free. This means that a small spacelike circle orthogonal to the congruence direction, is propagated along the congruence to remain a circle - rather than to be distorted into an ellipse. We may think of this shear-free condition as containing the main essence of the complex structure of $\mathbb{T}$. The remaining part has to do with the complex structure of the fibres in the projection $\mathbb{T}-\{0\} \rightarrow \mathbb{PT}$, i.e., of the 1-dimensional complex vector space of multipliers $Z^\alpha \mapsto \lambda Z^\alpha$, i.e., of $\pi_{A'} \mapsto \lambda\pi_{A'}$.

If we pass now to the curved case N, we see immediately that there is a serious problem. It is again the conformal curvature of the space-time that causes the difficulty. A congruence of null geodesics which starts out shear-free and which then enters a region of conformal curvature, will (except in very special circumstances) emerge on the other side with non-zero shear. This is in clear contrast with the case of rotation, which was defined above in terms of the symplectic form $\underset{\sim}{\omega}$. A non-rotating congruence of null geodesics remains non-rotating whatever kind of curvature it passes through. (In fact a non-rotating null congruence consists of the generators of a family of null hypersurfaces.) Thus, while the symplectic structure of twistor space may be regarded as invariant under the effects of curvature in M, the complex structure is 'shifted' as the null geodesics pass through a region of conformal curvature.

One point of view which was adopted in an earlier article$^{(32)}$ is that the complex structure of twistor space is relevant only to the quantum structure of nature. On the classical level one might be forced to consider the non-holomorphic pairs of quantities Z^{α}, $\bar{Z}_{\alpha}$ while on the quantum level it would be the 'holomorphic' pairs Z^{α}, $-\partial/\partial Z^{\alpha}$ which would be relevant (the operation of complex conjugation now playing no role). My present view of the situation is not greatly changed from this, but I do feel that this earlier attitude was perhaps unduly pessimistic as to the applicability of holomorphic twistor theory on the classical level - especially where solutions of Einstein's equations are involved. The reason I feel more hopeful now is that there have been several developments concerning the application of twistor-holomorphic ideas within classical curved space-times.

In the first place, it is now possible to make much more precise the rather vague concept of the 'shifting' of complex structure induced on N, by the presence of conformal curvature. The idea is first to show how a 'twistor space' $T(S)$ with a well-defined complex structure may indeed be defined, but relative to some hypersurface (preferably spacelike) in the space-time M. As S is moved to a new location S' we get a new complex space $T(S')$. We have $N \subset T(S)$ (assuming M is globally hyperbolic and S a spacelike Cauchy surface) and also $N \subset T(S')$, so N acquires two distinct structures, one from each of the complex 4-manifolds $T(S)$ and $T(S')$. The elements of any such space $T(S)$ are called *hypersurface twistors*.

Let us suppose that M is an analytic space-time and that S is an analytic hypersurface in M. (It is to be hoped that eventually this condition of analyticity may prove to be inessential. However, for the time-being it provides the only clear way of proceeding.) We can construct complexifications $\mathbb{C}M$ and $\mathbb{C}S$ ($\subset \mathbb{C}M$) of these manifolds, although we do not necessarily expect to be able to move very 'far' from the real sections $M \subset \mathbb{C}M$, $S \subset \mathbb{C}S$ before singularities (in, say, the complexified metric) are encountered. We should think of $\mathbb{C}M$ and $\mathbb{C}S$ as just 'thickenings' of M and S, respectively. Suppose, for the moment, that M is *flat*. Then any twistor Z^{α} in M defines, and is defined by, a totally null plane $\mathbb{C}Z$ and a spinor $\pi_{A'}$ such that the tangent vectors to $\mathbb{C}Z$ have the form $\xi^{A}\pi^{A'}$. The spinor $\pi_{A'}$ is constant throughout $\mathbb{C}Z$, so in particular it is constant along the (complex) curve ζ of intersection of $\mathbb{C}Z$ with $\mathbb{C}S$. Let n^{a} be normal to $\mathbb{C}S$. Then each tangent t^{a} to ζ is orthogonal to n^{a} and of the form $\xi^{A}\pi^{A'}$. This is achieved by taking

$$t^{AA'} = n^{AB'}\pi_{B'}\pi^{A'} \qquad (5.13)$$

so the condition that $\pi_{A'}$ be constant along ζ is

$$n^{AB'}\ \pi_{B'}\ \pi^{A'}\ \nabla_{AA'}\ \pi_{C'} = 0 \tag{5.14}$$

We can regard (5.14) as not only stating the constancy of $\pi_{A'}$ along ζ but also as giving a differential equation for ζ itself (since, given n^a, the direction of ζ is defined by (5.13)). Thus, the various solutions of (5.14) along $\mathbb{C}\,S$ define the various twistors in M which actually meet $\mathbb{C}\,S$. (The considerations here are for the most part local ones, so I shall not concern myself with the twistors which do not meet $\mathbb{C}\,S$.) Indeed, precisely the same construction can be given whether M is flat or not. The space of solutions of (5.14) on $\mathbb{C}\,S$ is the required hypersurface twistor space $T(S)$ and each solution is called a hypersurface twistor.

Note that (5.14) contains no reference to the operation of complex conjugation. It follows that $T(S)$ (assuming that it is a manifold at all - and in general it will be) must be canonically a complex manifold. Also it admits an operation of multiplication by complex numbers; thus if $Z^\bullet \in T(S)$ corresponds to $\pi_{A'}$ along ζ satisfying (5.14), then $\rho Z^\bullet \in T(S)$ corresponds to $\rho\pi_{A'}$ along the same ζ $(\rho \in \mathbb{C})$. Thus we can also consider a projective hypersurface twistor space $\mathbb{P}\,T(S)$ which, being the space of equivalence classes of proportional hypersurface twistors, may be regarded as the space of curves ζ. It should be remembered that there is also to be one special element of $T(S)$, called the zero element $0^\bullet$ which does not correspond to an element of $\mathbb{P}\,T(S)$.

We define $0Z^\bullet = 0^\bullet$ for each $Z^\bullet \in T(S)$; but otherwise we do not allow $\pi_{A'} = 0$. An element of $T(S)$ or of $\mathbb{P}\,T(S)$ is called null if the curve ζ contains a real point (point of S). Provided S is spacelike and the 'thickening' $\mathbb{C}S$ of S is sufficiently 'thin', none of the curves ζ will contain more than one real point. The system $\mathbb{P}\,N(S)$ of such curves containing a real point will be 5-real-dimensional; the corresponding system $N(S)$ of ζ's with $\pi_{A'}$'s attached is 7-real-dimensional. The whole space $T(S)$ is 4-complex- (or 8-real-) dimensional and $\mathbb{P}\,T(S)$ is 3-complex- (or 6-real-) dimensional.

The space $N(S)$ may be canonically identified with the space N of global null twistors (assuming M is globally hyperbolic and S a spacelike Cauchy hypersurface for the strict statement - otherwise some suitable modification is required). This is because the $\pi_{A'}$ at the real point of ζ (when $Z^\bullet \in N(S)$) defines a null geodesic in M passing through that point in the uniquely defined real null direction given by $p^{AA'} = \bar{\pi}^A \pi^{A'}$. Propagating $\pi_{A'}$ parallelly along this null geodesic we define a unique element of N as required. Conversely, each element of N defines a null geodesic and a $\pi_{A'}$. The null geodesic has a unique point of intersection with S through which passes a unique ζ-curve in $\mathbb{C}\,S$ uniquely defined by $\pi_{A'}$ at that point. This establishes the canonical identification of $N(S)$ with N. However $N\,(S)$ also inherits some structure from the complex manifold $T(S)$ in which it is imbedded. We see, in fact, that the hypersurface S supplies just that sought-for structure which was missing in the definition of N, namely that some non-null twistors should be provided which, when added to the space N, give rise to a 4-complex-dimensional manifold. This structure is, however, highly dependent on the particular choice of S in a general space-time. If S is moved to a new location S', the resulting structure for N will change - giving a clear realization to the 'shifting' of complex structure that was alluded to earlier.

Only in the case of a conformally flat space-time will this 'shifting' of structure be entirely absent. To see this, it will be helpful to examine the analogue of the Kerr theorem for hypersurface twistors. Since we have a well defined holomorphic operation $Z^{\bullet} \mapsto \rho Z^{\bullet}$, we can talk about homogeneous holomorphic functions on $T(S)$. We therefore can consider the loci on $\mathbb{C}T(S)$ defined by the vanishing of such homogeneous functions. Each of these loci corresponds to a congruence of ζ-curves on S; the ζ-curves of this congruence which contain real points (giving elements of $\mathbb{P}N(S)$, i.e., of $\mathbb{P}N$) define a congruence of null geodesics in M. Such a congruence of null geodesics is characterized by the fact that it is shear-free at its intersection with S. To see this we examine (5.14), selecting a particular choice of $\pi_{A'}$ for each null geodesic of the congruence. We now have $\pi_{A'}$ given as a function of position on S. Extending this selection of $\pi_{A'}$ to apply also to points of $\mathbb{C}S$ we obtain $\pi_{A'}$ as a spinor field on $\mathbb{C}S$. (The possibility of doing this requires the holomorphic nature of the function on $T(S)$.) Since these $\pi_{A'}$'s are associated with a congruence of ζ-curves we must have (5.14) holding, except for the fact that our arbitrary selection of one $\pi_{A'}$ (out of its proportionality class) per point of the curve may not agree with the scaling for $\pi_{A'}$ demanded by (5.14). Thus we have the weaker equation

$$n^{AB'}\pi_{B'}(\pi^{A'}\ \pi^{C'}\ \nabla_{AA'}\ \pi_{C'}) = 0 \tag{5.15}$$

holding on S (the extra $\pi^{C'}$ being inserted to kill the term arising from the derivative of the proportionality factor for $\pi_{C'}$). Equation (5.15) must now also hold for the $\pi_{A'}$-spinor field restricted to S. The portion of the expression in parenthesis will be recognized(7),(20) as the quantity whose vanishing signifies 'shear-freeness' for a congruence of null geodesics with tangent vector $p^{AA'} = \bar{\pi}^{A}\ \pi^{A'}$. The extra term $n^{AB'}\ \pi_{B'}$ really plays no role, since if we extend the spinor field on S to a spinor field throughout M, such that the p^{a}-field is tangent to a congruence of geodesics, we have that (5.15) holds if $n^{AB'}\ \pi_{B'}$ is replaced by $\bar{\pi}^{A}$ in any case. Thus, (5.15) implies (because of the 2-dimensionality of spin-space) that the same equation also holds on S if the term $n^{AB'}\ \pi_{B'}$ is omitted altogether. This establishes the generalized Kerr theorem for a hypersurface S.

Having this interpretation for the complex structure of $T(S)$ we see that it will indeed 'shift' as the surface S is varied unless M is conformally flat. The reason is, as was given earlier, that conformal curvature will cause a congruence of null geodesics which is initially shear-free generally to become shearing further along the congruence (Sachs equation)(7). It follows that a complex function on N, which is the restriction of some holomorphic function on $T(S)$, will not in general be the restriction of any holomorphic function on $T(S')$ where S' is a distinct hypersurface from S. On the other hand, when M is conformally flat, the holomorphic structure of N is the same whichever $T(S)$ is used.

We can also define the 'dual' hypersurface twistor space $T^*(S)$ by an exactly analogous construction to that of $T(S)$, but where primed and unprimed indices are interchanged in (5.14):

$$n^{BA'}\tilde{\pi}_B\tilde{\pi}^A\nabla_{AA'}\tilde{\pi}_C = 0 \quad . \tag{5.16}$$

This equation defines a unique complex curve $\tilde{\zeta}$ in $\mathbb{C}S$ through each point of $\mathbb{C}S$ and for each choice of $\tilde{\pi}_A$ at that point, the tangent direction to the curve being that of $n^{BA'}\tilde{\pi}_B\tilde{\pi}^A$. Such a curve $\tilde{\zeta}$, with $\tilde{\pi}_A$ propagated along it according to (5.16), defines a dual hypersurface twistor $\tilde{Z}_{\bullet} \in T^*(S)$. The use of a 'tilde' here is to indicate that $\tilde{\pi}_A$, $\tilde{\zeta}$ and $\tilde{Z}_{\bullet}$ are independent quantities from $\pi_{A'}$, ζ and $Z^{\bullet}$, but that the special case $\tilde{\pi}_A = \bar{\pi}_{A'}$, $\tilde{\zeta} = \bar{\zeta}$ defines us the complex conjugate $\bar{Z}_{\bullet}$ of the hypersurface twistor $Z^{\bullet}$. The space $T^*(S)$ has a complex structure just as $T(S)$ has. But the map from $T(S)$ to $T^*(S)$ defined by $Z^{\bullet} \mapsto \bar{Z}_{\bullet}$ is anti-holomorphic.

We can define a relation of incidence between $Z^{\bullet} \in T(S)$ and $\tilde{Z}_{\bullet} \in T^*(S)$, to be written

$$Z^{\bullet}\tilde{Z}_{\bullet} = 0 \quad , \tag{5.17}$$

by the condition that ζ and $\tilde{\zeta}$ have a point ($\in$ $\mathbb{C}S$) in common. The special case $Z^{\bullet}\bar{Z}_{\bullet} = 0$ is the defining relation for $Z^{\bullet} \in N(S)$. This is because the operation of complex conjugation applied to $\mathbb{C}M$ and $\mathbb{C}S$

sends ζ into $\bar{\zeta}$ and must send any (necessarily unique) intersection point of ζ with $\bar{\zeta}$ into itself. Such a point must therefore be real and the conclusion follows. When M is flat (or conformally flat), (5.17) agrees (at least for twistors for which $\mathbb{C}\,Z$ passes 'near enough' to S) with the standard relation $Z^{\alpha}\tilde{Z}_{\alpha} = 0$. As yet, no satisfactory definition of scalar product has emerged, in the general case, which would extend the relation (5.17) to give a value to '$Z^{\cdot}\tilde{Z}_{\cdot}$' in the case of arbitrary elements $Z^{\cdot} \in T(S)$, $\tilde{Z}_{\cdot} \in T^{*}(S)$.

Having a suitable complex space $T(S)$ we can actually construct an analogue of (3.10) for the generation of zero rest-mass fields from holomorphic functions of twistors. However this seems, at present, to be of limited value because the contour integral can (apparently) only be performed for space-time points on the hypersurface S (or $\mathbb{C}S$). Thus it seems that only 'constraint equation' type relations can be automatically satisfied in this way, and not the propagation of fields away from S. On the other hand, the situation is considerably better in the case of the two-variable twistor functions. A generalization to M of (3.39) does appear to be possible, for generating solutions of (3.41). We are concerned here with holomorphic functions $F(W_{\cdot}, Z^{\cdot})$ where $W_{\cdot} \in T^{*}(S)$ and $Z^{\cdot} \in T(S)$, that is, with holomorphic maps

$$F : T^{*}(S) \times T(S) \rightarrow \mathbb{C} \quad . \tag{5.18}$$

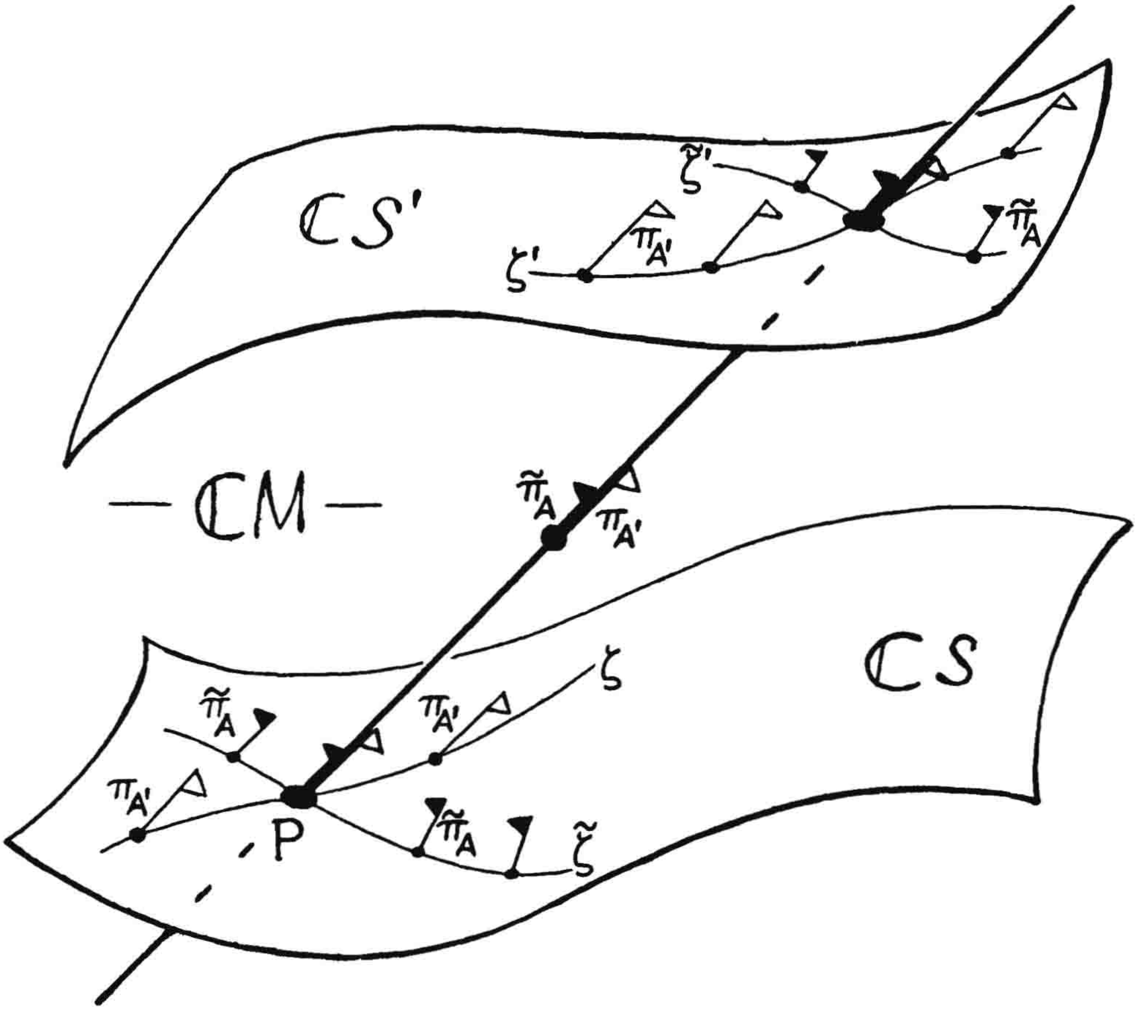
Since we have a well-defined scalar multiplication on each of $T^{*}(S)$ and

$T(S)$ there is no problem about stipulating the homogeneity requirements for F. Nor is there any real problem about interpreting the contour integral (3.39) for space-time points on S or $\mathbb{C}S$. What is required is a method of dealing with points not on $\mathbb{C}S$.

Let us examine the space $T^*(S) \times T(S)$ on which F is defined. To begin with, it is 8-complex-dimensional, having a well-defined complex structure. Furthermore, it contains a 7-complex-dimensional subspace

$$N(S) \subset T^*(S) \times T(S) \tag{5.19}$$

consisting of those pairs $(\tilde{Z}_{\cdot}, Z^{\cdot})$ for which the incidence relation $Z^{\cdot}\tilde{Z}_{\cdot} = 0$ holds. The reason for the notation in (5.19) is as follows. Whenever $Z^{\cdot}\tilde{Z}_{\cdot} = 0$, we have a point $P \in \mathbb{C}S$ on both curves $\zeta, \tilde{\zeta}$ and a pair of spinors $\pi_{A'}, \tilde{\pi}_A$ at P. We therefore have a unique complex null geodesic in $\mathbb{C}M$ through P in the direction of the complex null vector $\tilde{\pi}_A \pi_{A'}$. The spinors $\tilde{\pi}_A$ and $\pi_{A'}$ can then be carried up this null geodesic by parallel propagation (with respect to the complex Riemannian connection defined by the complex metric of $\mathbb{C}M$), thus freeing the description from the particular hypersurface S. Any other complex hypersurface $\mathbb{C}S'$ which intersects this null geodesic could be used and the description referred, instead, to $T^*(S') \times T(S')$. (See Fig. 6.8). When $\tilde{Z}_{\cdot} = \bar{Z}_{\cdot}$ we have $\tilde{\zeta} = \bar{\zeta}$, $P \in S$, $\tilde{\pi}_A = \bar{\pi}_A$, the above null geodesic becoming the real null geodesic Z and the description agreeing with that for a global null twistor ($\in N$). The more general case, described above, when $\tilde{Z}_{\cdot}$ need not coincide with $\bar{Z}_{\cdot}$ gives an



$\mathbb{C}S'$
$\tilde{\zeta}'$
ζ'
$\pi_{A'}$
$\tilde{\pi}_A$
$\mathbb{C}M$
$\tilde{\pi}_A$
$\pi_{A'}$
$\mathbb{C}S$
ζ
$\tilde{\pi}_A$
$\pi_{A'}$
$\pi_{A'}$
$\tilde{\pi}_A$
P
$\tilde{\zeta}$

Fig. 6.8

element of a suitably complexified space $\mathbb{C}N$. Not every element of $\mathbb{C}N$ defines a null geodesic which actually meets $\mathbb{C}S$. Those which do can be identified with elements of $\mathbb{C}N(S)$, so with this identification made, we should have

$$\mathbb{C}N(S) \subset \mathbb{C}N \tag{5.20}$$

(although it may be preferred, for logical reasons, not to make this identification but refer instead to an injective map $\mathbb{C}N(S) \to \mathbb{C}N$).

Suppose, now, that we have two analytic Cauchy hypersurfaces S, S' for M. Then we have

$$N \subset \mathbb{C}N(S,S') \subset \mathbb{C}N \tag{5.21}$$

where (adopting the identification (5.20)),

$$\mathbb{C}N(S,S') = \mathbb{C}N(S) \cap \mathbb{C}N(S'). \tag{5.22}$$

The complex space $\mathbb{C}N(S,S')$, being regarded both as a subspace of $T^*(S) \times T(S)$ and of $T^*(S') \times T(S')$ can thus be described equally, in terms of hypersurface twistor pairs $(W_{\cdot}, Z^{\cdot})$ with $W_{\cdot} \in T^*(S)$, $Z^{\cdot} \in T(S)$ or pairs $(U_{\cdot}, X^{\cdot})$ with $U_{\cdot} \in T^*(S')$, $X^{\cdot} \in T(S)$. This is not to say that $W_{\cdot}$ may in any way be identified with $U_{\cdot}$, or $Z^{\cdot}$ with $X^{\cdot}$ (except if M is conformally flat). It is just that the pair $(W_{\cdot}, Z^{\cdot})$ may be relabelled as the pair $(U_{\cdot}, X^{\cdot})$. (If $W_{\cdot}$ is varied but $Z^{\cdot}$ held fixed then, in general, both $U_{\cdot}$ and $X^{\cdot}$ will vary, for example.)

It should now be clear (in principle at least) how the function F of (5.18) is to be interpreted, as regards its employment in contour integrals like (3.39), when the space-time point in question does not lie on $\mathbb{C}\,\mathcal{S}$. We simply reinterpret F as a function of $U_{\bullet}$ and $X^{\bullet}$ where the hypersurface $\mathcal{S}'$ (or $\mathbb{C}\,\mathcal{S}'$) is chosen to pass through this point. Indeed, we may prefer to regard F as defined on $\mathbb{C}\,\mathcal{N}$ and dispense with the hypersurface twistor description altogether! However, it seems likely that to do so would be to forfiet some valuable potential insights. (It should be pointed out, in this connection, that the values of F on $\mathbb{C}\,\mathcal{N}$ are not, in fact, sufficient to define the function F of (5.18), since this also requires pairs $(W_{\bullet}, Z^{\bullet})$ for which $Z^{\bullet} W_{\bullet} \neq 0$.)

Since $\mathcal{M}$ is analytic, the forms $\underline{\phi}$ and $\underline{\omega}$, belonging to the invariant structure of $\mathcal{N}$, can be 'complexified' so as to refer to $\mathbb{C}\mathcal{N}$. It is helpful also to consider the complex conjugate $\bar{\underline{\phi}}$ of $\underline{\phi}$ as something independent of $\underline{\phi}$ even though $\bar{\underline{\phi}} = \underline{\phi}$ on $\mathcal{N}$. This gives us another 1-form $\tilde{\underline{\phi}}$ (the complexified $\bar{\underline{\phi}}$) defined on $\mathbb{C}\,\mathcal{N}$ which takes the same values as $\underline{\phi}$ on $\mathbb{C}\,\mathcal{N}$, but whose functional form is different. We can refer these forms either to $\mathcal{S}$ or to $\mathcal{S}'$ and write

$$
\begin{aligned}
\underline{\omega} &= i\, dZ^{\bullet} \wedge dW_{\bullet} = i\, dX^{\bullet} \wedge dU_{\bullet} \\
\underline{\phi} &= i\, Z^{\bullet}\, dW_{\bullet} = i\, X^{\bullet}\, dU_{\bullet} \\
\tilde{\underline{\phi}} &= -\, i\, W_{\bullet}\, dZ^{\bullet} = -\, i\, U_{\bullet}\, dX^{\bullet}
\end{aligned}
\qquad (5.23)
$$

although some detailed interpretation of the meanings of these symbols is required. The first of the relations (5.23) suggests that the passage from a $(W_{\bullet}, Z^{\bullet})$ description to a $(U_{\bullet}, X^{\bullet})$ description should be regarded as a canonical transformation. For this to be strictly true, however, we would need to factor out $\mathbb{C}\mathcal{N}$ by the equivalence relation

$$(W_{\bullet}, Z^{\bullet}) \equiv (\rho^{-1} W_{\bullet}, \rho Z^{\bullet}); \quad (U_{\bullet}, X^{\bullet}) \equiv (\sigma^{-1} U_{\bullet}, \sigma X^{\bullet}) \qquad (5.24)$$

which now gives us a 6-complex-dimensional space (a complexification of the space of scaled null geodesics in $\mathcal{N}$, cf. (5.12) with $\rho = e^{i\theta}$) on which $\underline{\omega}$ is now a non-degenerate closed 2-form defining a (complex) symplectic structure. In any case, the framework of Hamiltonian techniques and generating functions are at hand and should yield insights into the nature of the transformations between hypersurface twistors defined on different hypersurfaces. In particular, it may be pointed out that if a generating function $g(W_{\bullet}, X^{\bullet})$ is chosen which is homogeneous of degrees (1,1), then - formally (at least) - the equations

$$Z^{\bullet} = \partial g/\partial W_{\bullet}, \qquad U_{\bullet} = \partial g/\partial X^{\bullet} \qquad (5.25)$$

lead to satisfaction of the second and third relations (5.23) in addition to the first one.

An obvious drawback to the notion of hypersurface twistor is its dependence on the location of an arbitrary hypersurface $\mathcal{S}$. Another serious drawback is the absence, so far at least, of a general notion of scalar product $Z^{\bullet} W_{\bullet}$. Both of these drawbacks can be overcome in the case of an asymptotically flat space-time, if we specialize S by allowing it to recede into the remote (null) past or future(4),(7),(28) to become one of the two null conformal infinities $\mathscr{I}^-$ or $\mathscr{I}^+$. For this, it is unnecessary to take a limit because the construction can simply be applied to $\mathscr{I}^\pm$ directly. However, one must verify that the construction is invariant under a conformal rescaling of the metric, since such a rescaling has first to be made in order that $\mathscr{I}^\pm$ can be treated as finite hypersurfaces. Indeed, all that is required is the observation that (5.14) remains valid under (2.16), because of (2.17), where $\hat{\pi}_{A'} = \pi_{A'}$ in accordance with (2.19) on $\omega^A = 0$. The entire theory of hypersurface twistors is thus independent of the conformal scale.

In much of the discussion of hypersurface twistors it was remarked that the construction works best when S is taken to be a spacelike Cauchy hypersurface. Null hypersurfaces in general could lead to some problems, although the construction still applies locally in the null case. The particular null hypersurfaces $\mathscr{I}^\pm$, on the other hand, have some very simple features, so it may reasonably be said that in this case the construction works better than it would for a general spacelike hypersurface. Furthermore, although $\mathscr{I}^\pm$ are not, strictly, Cauchy hypersurfaces for M, they do (for asymptotically simple M) serve the same essential purpose of intersecting every null geodesic in M.

To discuss the spaces $T(\mathscr{I}^+)$ and $T^*(\mathscr{I}^+)$ in any detail it will be necessary first to have some picture of the structure of $\mathbb{C}\mathscr{I}^+$. (The discussion for $\mathscr{I}^-$ would be similar - although a lot simpler in the case of retarded fields since it is the presence of gravitational radiation crossing $\mathcal{I}^{\pm}$ that causes all the difficulties.) Now the normal n^a to any null hypersurface has the form

$$n^{AA'} = \iota^{A}\bar{\iota}^{A'} \tag{5.26}$$

The complex tangent vectors are thus linear combinations of complex null vectors of the two types

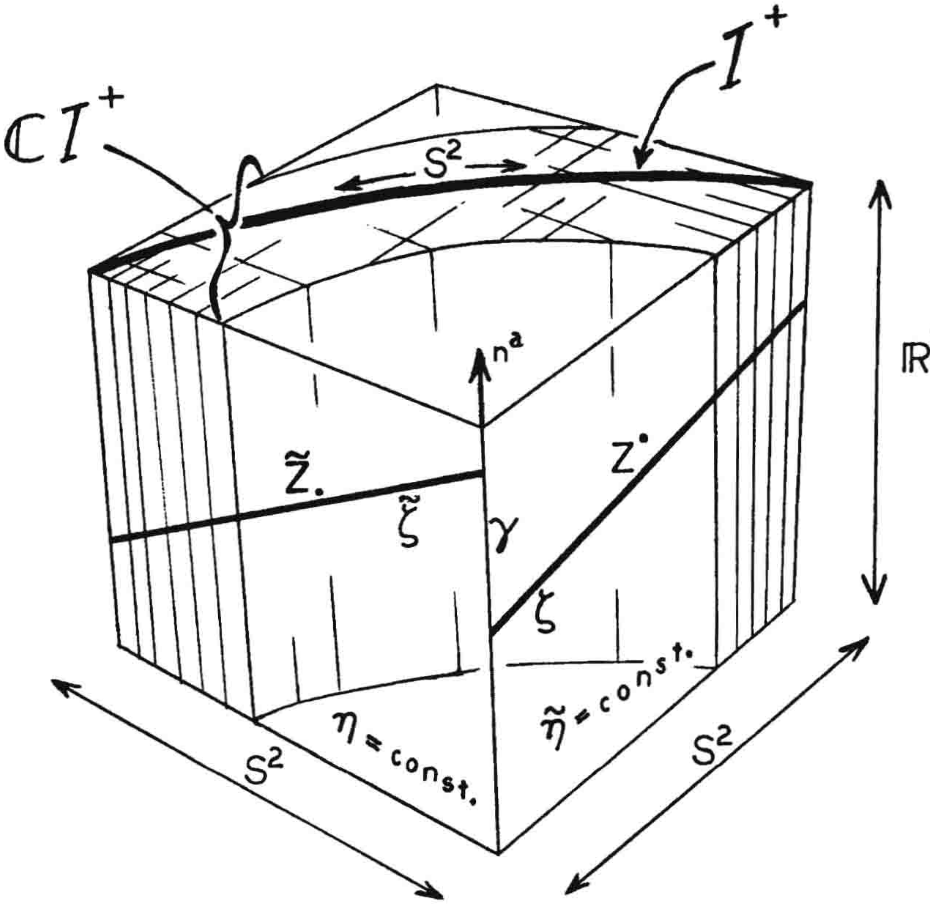
$$\iota^{A}\lambda^{A'} \quad , \qquad \mu^{A}\bar{\iota}^{A'} \tag{5.27}$$

In the case of $\mathbb{C}\,\mathcal{I}^{+}$, each of (5.27) is tangent to a totally null complex plane whose other tangents have all the same form (including (5.26) as a special case). In fact, this property is a consequence of the vanishing of the shear of the generators (integral curves of (5.26)) of $\mathcal{I}^{+}$ - although I shall not show this here.

It is common practice to choose coordinates and conformal scale for $\mathcal{I}^{+}$ so that its intrinsic metric takes the form

$$ds^2 = -\frac{4d\eta d\bar{\eta}}{1+\eta\bar{\eta}} + 0.\ du^2 \tag{5.28}$$

the complex coordinate η being related to the usual spherical polars by $\eta = e^{i\phi}\cot\frac{1}{2}\theta$ and u being a Bondi-type (say), retarded time(44),(45),(4). The complex coordinate η ranges over the whole Riemann sphere ($\eta = \infty$ included) giving $\mathcal{I}^{+}$ the topology $S^2 \times \mathbb{R}$. To complexify $\mathcal{I}^{+}$, we must allow η to become 'freed' from η and u to become a complex parameter. Thus we write $\tilde{\eta}$ for $\bar{\eta}$. The whole range of coordinate variables would give a space with topology $S^2 \times S^2 \times \mathbb{R}^2$. The totally null planes mentioned above are now given by η = const. and $\tilde{\eta}$ = const., respectively, so as far as the coordinate range is concerned, each such plane would have topology $S^2 \times \mathbb{R}^2$. However, in general one can only expect $\mathbb{C}\mathscr{I}^+$ to cover some open subset of this coordinate range, which includes the 'diagonal' $S^2 \times \mathbb{R}^1$ given by $\mathscr{I}^+$ itself (see Fig. 6.9). Even in the case when M is Minkowski space $\mathbb{M}$ the topology of $\mathbb{C}\mathscr{I}^+$ is not quite $S^2 \times S^2 \times \mathbb{R}^2$ but a (twisted) $\mathbb{R}^2$ bundle over $S^2 \times S^2$. (This arises because the complexified u is not a



$\mathbb{C}\mathscr{I}^+$
$\mathscr{I}^+$
S^2
$\mathbb{R}^2$
n^a
$\tilde{Z}.$
$Z^{\cdot}$
$\tilde{\zeta}$
γ
ζ
$\tilde{\eta}$ = const.
η = const.
S^2
S^2

Fig. 6.9

good coordinate over the whole of $\mathbb{C}\mathcal{I}^+$.) The η = const. and $\tilde{\eta}$ = const. totally null planes are likewise topologically twisted.

Let us now see how the hypersurface twistor construction applies to $\mathbb{C}\mathcal{I}^+$. Equation (5.14) now becomes (with (5.26))

$$\iota^A \pi^{A'} \nabla_{AA'} \pi_{B'} = 0 \tag{5.29}$$

assuming that $\pi_{B'}$ is not proportional to $\tilde{\iota}_{B'}$. It is possible$^{(28)}$ to choose ι^A and a new conformal scale so that

$$\nabla_{AA'} \iota_B = 0 \tag{5.30}$$

Combining this with (5.29) we see that the ζ-curves are now null geodesics on $\mathbb{C}\mathcal{I}^+$. (Recall that the concept of an unscaled null geodesic is in fact conformally invariant.) These null geodesics lie entirely in the $\tilde{\eta}$ = const. planes. In a similar way, equation (5.16) becomes

$$\tilde{\pi}^A \tilde{\iota}^{A'} \nabla_{AA'} \tilde{\pi}_B = 0 \tag{5.31}$$

showing that the $\tilde{\zeta}$ curves are geodesics lying in the η = const. planes. An asymptotic twistor, therefore, is an element $Z^{\cdot} \in T(\mathcal{I}^+)$, corresponding to a null geodesic ζ in $\mathbb{C}\mathcal{I}^+$ with tangent vector $\iota^A \pi^{A'}$, where $\pi_{A'}$ is parallelly propagated along ζ (ζ lying in some $\tilde{\eta}$ = const.); an element $\tilde{Z} \in T^*(\mathcal{I}^+)$ corresponds to a null geodesic $\tilde{\zeta}$ in $\mathbb{C}\mathcal{I}^+$ with tangent vector $\tilde{\pi}^{A}\,\tilde{\iota}^{A'}$, where $\tilde{\pi}_{A}$ is parallel along $\tilde{\zeta}$ (with $\tilde{\zeta}$ in some η = const.). This is assuming that $\pi_{A'}\,\tilde{\iota}^{A'} \neq 0 \neq \iota^{A}\,\tilde{\pi}_{A}$. There are also elements of $\mathcal{I}(\mathcal{I}^{+})$ and $T^{*}(\mathcal{I}^{+})$ corresponding to the directions $\iota^{A}\,\tilde{\iota}^{A'}$, but their definition is more complicated and will not be entered into here. (See ref. (19) for a discussion of this and other aspects of the nature of $T(\mathcal{I}^{+})$ and $T^{*}(\mathcal{I}^{+})$.) These special asymptotic twistors 'entirely on $\mathcal{I}^{+}$' are the analogues of flat space twistors having standard representations $(\omega^{A},0)$ or $(0,\tilde{\omega}^{A'})$, i.e., being of the form $I^{\alpha\beta}\,K_{\beta}$ or $I_{\alpha\beta}\,L^{\beta}$.

Let us see how the scalar product $Z^{\bullet}\,\tilde{Z}_{\bullet}$ of an element $Z^{\bullet} \in T(\mathcal{I}^{+})$ and an element $Z_{\bullet} \in T^{*}(\mathcal{I}^{+})$ is to be formed. In unfavourable circumstances, the situation is as depicted in Fig. 6.9, where the intersection γ of the $\tilde{\eta}$ = const. plane containing ζ with the η = const. plane containing $\tilde{\zeta}$ lies outside the region covered by $\mathbb{C}\,\mathcal{I}^{+}$. In such circumstances, the scalar product $Z^{\bullet}\,\tilde{Z}_{\bullet}$ is deemed to have become singular (unless some alternative definition has been found which analytically extends the definition from the regular regions). Let us suppose, instead, that the circumstances are favourable and that γ not only intersects $\mathbb{C}\,\mathcal{I}^{+}$ but intersects it in a way that a continuous path α exists in $\zeta \cup \gamma \cup \tilde{\zeta}$ connecting a point Q in ζ to a point $\tilde{Q}$ in $\tilde{\zeta}$. (It would be convenient, also, to assume that this path is unique up to homotopy - since this would ensure the uniqueness of the definition.) The scalar product is now obtained by giving a local twistor description of $Z^{\bullet}$, namely $(0,\pi_{A'})$, at the point Q and carrying this along α by local twistor transport to the point $\tilde{Q}$, at which point the resulting local twistor has its scalar product formed in the normal way with the local twistor description $(\tilde{\pi}_A, 0)$ at $\tilde{Q}$ of $\tilde{Z}_\bullet$. The properties of local twistor transport (cf. (5.3), (5.6)) ensure that this definition is independent of the precise location of Q and $\tilde{Q}$ – we may as well choose them as the respective intersections of ζ and $\tilde{\zeta}$ with γ – and that the same answer is obtained if we proceed in the reverse order from $\tilde{Q}$ to Q.

We now have $Z^\bullet \tilde{Z}_\bullet$ defined as a holomorphic function on some open subset of $T(I^+) \times T^*(I^+)$. This open subset contains $\mathbb{C}N(I^+)$, where the definition $Z^\bullet \tilde{Z}_\bullet = 0$ agrees with the one given earlier in (5.17). Since we now have $Z^\bullet \tilde{Z}_\bullet$ defined in a neighbourhood of $\mathbb{C}N(I^+)$ we can take derivatives (separately) with respect to $Z^\bullet$ and $\tilde{Z}_\bullet$. This enables us to obtain the differential forms (5.23) in this way and to assign a more precise meaning to the notation given there. (We might, for example, take $W_\bullet \in T^*(I^+)$, $Z^\bullet \in T(I^+)$, $U_\bullet \in T^*(I^-)$, $X^\bullet \in T(I^-)$.) The quantity obtained by differentiating $Z^\bullet \tilde{Z}_\bullet$ once independently with respect to each variable $Z^\bullet$, $\tilde{Z}_\bullet$ and then setting $\tilde{Z}_\bullet = \bar{Z}_\bullet$ defines a Hermitian metric tensor for $T(I^+)$ of signature $(+ + - -)$ – at least at those points of $T(I^+)$, near enough to $N(I^+)$, for which $Z^\bullet \tilde{Z}_\bullet$ is unambiguously defined – and gives (this portion of) $T(I^+)$ the structure of a Kähler manifold(43). The form $\underline{\omega} = i\, dZ^\bullet \wedge d\bar{Z}_\bullet$ of (5.23) is the Kähler 2-form.

There is much that can be said about the details of this Kähler structure; in particular, the Kähler curvature can be related to the gravitational radiation(19). I shall not enter into the details of any of this here. The important thing to realize is that $T(I^+)$ has a rich structure which can be related in detail to physically meaningful asymptotic properties of the space-time. It seems that this structure is adequate for the definition of the basic mathematical operations required for a twistor scattering theory - although, much remains to be learned as to the most satisfactory method of handling this structure. Furthermore, it should be emphasized that in the presence of outgoing gravitational radiation, the asymptotic twistor space $T(I^+)$ is curved, so that its structure differs in an essential way from that of the standard twistor space $\mathbb{T}$. When the space-time is stationary then $T(I^+)$ and $\mathbb{T}$ has isomorphic structures. (Absence of outgoing radiation may be sufficient for this.)

A final remark should be made concerning the global properties of the complex manifolds $T(S)$, $\mathbb{C}N$ and $T(I^+)$. It is a remarkable fact that complex manifolds are sometimes able to 'store' a great deal of information in their global structure, even though locally the only invariant quantity is the dimension of the manifold. It turns out, in fact, that in the generic case, the global complex structure of $T(S)$ is sufficient to enable part of the conformal geometry of $\mathbb{C}M$ at the points of $\mathbb{C}S$ to be reconstructed. Moreover, from the global complex structure of $\mathbb{C}N$ the conformal geometry of $\mathbb{C}M$ can, in general, be reconstructed. This requires that $\mathbb{C}N$ actually refer to the space of complex null geodesics in $\mathbb{C}M$ and is not simply a (local) complexification of the space N . This enables $\mathbb{C}N$ to acquire a necessary degree of compactness in certain directions. The global complex structure of $T(I^+)$ apparently carries information about the outgoing gravitational radiation field. In each case the global complex structure is different from that which could be obtained by taking an open subset of the complex spaces arising from Minkowski space by the corresponding constructions.

This can be illustrated by considering compact holomorphic curves in $\mathbb{PT}$, $\mathbb{P}\,T(S)$ and $\mathbb{P}\,T(I^{+})$. In the case of $\mathbb{PT}$ we have a particular complex 4-parameter infinity of such curves, namely the projective lines in $\mathbb{PT}$, which represent the points of $\mathbb{CM}$ (see §2). However for $\mathbb{P}\,T(S)$ this family becomes modified, a 3-parameter subsystem representing the points only of $\mathbb{C}\mathcal{S}$ arising in this way. The other points of $\mathbb{C}M$ cannot be so represented by compact holomorphic curves in the general case. As for $\mathbb{P}\,T(I^{+})$ it seems that we again get a complex 4-parameter family of holomorphic curves (of the same homology class as before - and assuming that the outgoing gravitational radiation field is not too 'rough'). But now these curves do not in general correspond to actual points of $\mathbb{C}M$, but define(46) the points of a new space H (colloquially referred to as 'Heaven'), which may be thought of as the space of complex slices of $\mathbb{C}I^{+}$ for which a certain measure of shear vanishes. This interpretation arises when we adapt the generalized Kerr theorem referred to earlier (cf. (5.15)) to the case $S = I^{+}$. (In this form, the theorem is due to E.T. Newman and co-workers.(47))

Newman has also shown(46) that H can be given a complex Riemannian structure in a natural way. The further geometry of H is presently under investigation.* It turns out the space H has the interesting property

CHAPTER 69

---

* It now emerges that H also satisfies Einstein's vacuum equations!

that its complex conformal curvature, when expanded in the form (3.8)

$$C_{AA'BB'CC'DD'} = \Psi_{ABCD}\,\varepsilon_{A'B'}\,\varepsilon_{C'D'} + \varepsilon_{AB}\,\varepsilon_{CD}\,\tilde{\Psi}_{A'B'C'D'} \;, \qquad (5.32)$$

satisfies $\Psi_{ABCD} = 0$, whereas in general $\tilde{\Psi}_{A'B'C'D'} \neq 0$. This can only happen because $\mathcal{H}$ is a complex Riemannian space.

6. OUTLOOK

In the preceding sections I have tried to indicate the essential scope of twistor theory as it now stands and to mention some of its details. There have been some omissions, the most important of which are the twistor Hamiltonian scattering theory (in which significant new contributions have been made by J. Moore, Z. Perjés and P. Tod), the explicit evaluation of many twistor diagrams (particularly some new work by G. Sparling) and some results (by A. Hodges) on massive free fields. Twistor theory remains a subject of active development. But there are a number of areas where new insights are clearly required. Some well-defined rule for generating twistor diagrams together with a suitable rule for the contour topology is needed. It may well be, in this connection, that a perturbation approach to scattering theory is inappropriate and what really is required is a single 'diagram' for generating an entire process. Such ideas are highly speculative at the present time although it is possible to attempt certain guesses. More insights are required as to the most natural way of breaking conformal invariance in twistor theory. Much more is also needed in the specific treatment of the electromagnetic and gravitational fields, some new ideas arising here being in an incomplete state.

In the case of gravity, one would wish to be able to present curved space-times somehow in terms of twistor holomorphic functions, a simple way of expressing satisfaction of Einstein's equations being clearly desirable. In fact there are some indications that such a programme may actually be realizable. Indeed, it is possible to specify space-time points in curved asymptotically flat (analytic) space-time in terms of holomorphic functions on asymptotic twistor space.* In this way, local considerations of space-time manifolds can be translated into statements about holomorphic functions. There appear also to be some suggestive analogies between this twistor approach to the study of curved space-times and the Souriau - Kostant theory of geometric quantization(48),(49). This encourages my own belief(32),(5) that twistors may reveal a hidden interrelationship between classical general relativity and quantum mechanics.

---

* Let f be holomorphic and homogeneous of degree + 2 on some suitable subset of $\mathcal{T}(\mathcal{I}^+)$. Then the pairs $(Z^\bullet, W_\bullet) \in \mathcal{T}(\mathcal{I}^+) \times \mathcal{T}^*(\mathcal{I}^+)$ for which $f(Z^\bullet) = 0$ and $W_\bullet = \partial f/\partial Z^\bullet$ are simultaneously satisfied define a subset of $\mathbb{P}\mathbb{C}\mathcal{N}$ representing the generators of a null hypersurface in $\mathbb{C}\mathcal{M}$. The problem is to select a system of f's for which these hypersurfaces may be construed as being the light cones of the points of a space-time $\mathcal{M}$ satisfying Einstein's equations.

Nature has, after all, chosen to weave her universe from these two constituents - and from others as yet largely unknown. Interrelationships must be present that we are unable now to perceive. And we are blinded not just by our lack of knowledge; our preconceived notions concerning space, time and quantum mechanics may be also partly to blame.

ACKNOWLEDGEMENTS

I am grateful to numerous colleagues and students for helping to supply many of the ideas I have been describing, some of whom have been explicitly referred to in the text. But there are many others; in particular, valuable suggestions have come from L.P. Hughston and especially C.J. Isham. Additional thanks are due to B.D. Bramson for making a great number of helpful criticisms concerning the manuscript. And I am indebted to Judith Daniels for suggesting improvements to a portion of the manuscript and more profoundly, for allowing me to air thoughts to her in their formative stages. In consequence, many thoughts formed which would otherwise not have.

REFERENCES

(1) C.F. von Weiszacker, *Brit. J. Phil. Sci.* (1973) 321; also, in *Quantum Theory and Beyond* (Ed. Ted. Bastin; Cambridge Univ. Press, 1971).

(2) R. Penrose, in *Quantum Theory and Beyond* (Ed. Ted Bastin; Cambridge Univ. Press, 1971).

(3) E. Cartan, *Ann. Sci. Ecole Normale Superieure* 31 (1914) 263.

(4) R. Penrose, in *Group Theory in Non-Linear Problems* (Ed. A.O. Barut; D. Reidel 1974)

(5) R. Penrose, in *Magic without Magic*, John Archibald Wheeler (ed. J.R. Klauder, W.H. Freeman & Co., San Francisco 1972).

(6) B.D. Bramson, *Physics Letters* 47A (1974) 431.

(7) R. Penrose, in *Battelle Rencontres* 1967 (Eds. C.M. DeWitt & J.A. Wheeler; Benjamin, New York 1968).

(8) J. Ehlers & A. Schild, *Comm. Math. Phys.* 32 (1973) 119.

(9) H. Bateman *Mathematical Analysis of Electrical and Optical Wave Motions* (Camb. Univ. Press 1915).

(10) I. Robinson & A. Trautman, *Proc. Roy. Soc.* (*Lond*) A265 (1962) 463.

(11) I. Robinson, *J. Math. Phys*. 2 (1961) 290.

(12) R.P. Kerr & A. Schild, *Proc. Symposia in Appl. Math*. 17 (1965) 199.

(13) I. Robinson, J.R. Robinson & J.D. Zund, *J. Math. Mech*. 18 (1969) 881.

(14) E.T. Newman & A.I. Janis, *J. Math. Phys*. 6 (1965) 915.

(15) E.T. Newman, *J. Math. Phys*. 14 (1973) 102, 774.

(16) L. Landau, in *Niels Bohr and the Development of Physics*, Pergamon, Lond, 1955.

(17) B.S. DeWitt, *Phys. Rev. Lett*. 13, (1964) 114.

(18) A. Salam & J. Strathdee, *Nuovo Cim. Lett*. 4 (1970) 101.

(19) R. Penrose & M.A.H. MacCallum, *Physics Reports* 6C (1973) 242.

(20) R. Penrose, *J. Math. Phys.* 8 (1967) 345.

(21) N.H. Kuiper, *Ann. Math* 50 (1949) 916.

(22) E. Grgin, Ph. D Thesis, Syracuse University (1966)

(23) J. Wess & B. Zumino, *Nuclear Phys*. B70, (1974) 39.

(24) J.G. Semple & L. Roth, *Algebraic Geometry* (Clarendon Press Oxford 1949).

(25) P.A.M. Dirac, *Proc. Roy. Soc. (Lond.)* A155 (1936) 447.

(26) M. Fierz & W. Pauli, *Proc. Roy. Soc. (Lond.)* A173 (1939) 211.

(27) M. Fierz, *Helv. Phys. Acta* 13 (1940) 45.

(28) R. Penrose, *Proc. Roy. Soc. (Lond.)* A284 (1965) 159.

(29) O. Laporte & G.E. Uhlenbeck, *Phys. Rev*. 37 (1931) 1380.

(30) R.K. Sachs & P. Bergmann, *Phys. Rev.* 112 (1958) 674.

(31) R. Penrose, *J. Math. Phys*. 10 (1969) 38.

(32) R. Penrose, *Int. J. Theor. Phys.* 1 (1968) 61.

(33) W. Rühl, in *Group Theory in Non-Linear Problems* (Ed. A.O. Barut; D. Reidel (1974)).

(34) M. Sato, *J. Fac. Sci. Univ. Tokyo*, Sect. I, 8 (1959/60) 139, 387.

(35) R. Penrose, in *Quantization of generally covariant field theories* (A.R.L. Tech. Doc. Rep. 63-56, P.G. Bergmann 1963).

(36) E.T. Newman & R. Penrose, *Proc. Roy. Soc.* (*Lond.*) A305 (1968) 175.

(37) A. Hodges, Preprint 1974.

(38) G.A.J. Sparling, Chapter 7, this volume.

(39) G.A.J. Sparling, Ph. D. Thesis, London University (1974).

(40) L. Gross, *J. Math. Phys*. 5 (1963) 687.

(41) P.A.M. Dirac, *The Principles of Quantum Mechanics* (Ox. Univ. Press, 1930).

(42) R.P. Geroch, *J. Math. Phys*. 11 (1970) 437.

(43) E. Nelson, *Tensor Analysis* (Princeton U. Press 1967).

(44) H. Bondi, A.W.K. Metzner & M.J.G. van der Burg, *Proc. Roy. Soc*. (Lond.) A269 (1962) 21.

(45) R.K. Sachs, *Proc. Roy. Soc. (Lond.)* A270 (1962) 103.

(46) E.T. Newman, report to Tel Aviv conference 1974.

(47) B. Aronson, R. Lind, J. Messmer & E.T. Newman, *J. Math. Phys*. 12 (1971) 2462.

(48) B. Kostant, in *Lectures in modern analysis and applications*, III, Springer Lecture Notes 170 (1970).

(49) J.M. Souriau, *Structures des systèmes dynamiques* (Dunod, 1970).

(50) Z. Perjés, *Twistor Variables of Relativistic Mechanics* (Preprint, Budapest, 1974).

Document generated by Anna’s Archive around 2023-2024 as part of the DuXiu collection (https://annas-blog.org/duxiu-exclusive.html).

Images have been losslessly embedded. Information about the original file can be found in PDF attachments. Some stats (more in the PDF attachments):

```
{
  "filename": "NDA3ODg1NDAuemlw",
  "filename_decoded": "40788540.zip",
  "filesize": 166490945,
  "md5": "72bdf4904495ecd16fb8e67fc3a1b7ba",
  "header_md5": "e25e4a4210345c6dd8feebe1fd42436e",
  "sha1": "badb81e3ca28b600a62d8962a2e2525ed18d2e7c",
  "sha256": "105c9a9dcfe3e53017ae050d207a66c755a65cb403457e24ce8972d315c85530",
  "crc32": 1122802796,
  "zip_password": "",
  "uncompressed_size": 207528862,
  "pdg_dir_name": "ROGER PENROSE COLLECTED WORKS VOLUME 2 1968-1975_40788540",
  "pdg_main_pages_found": 935,
  "pdg_main_pages_max": 935,
  "total_pages": 958,
  "total_pixels": 5888129736,
  "pdf_generation_missing_pages": false
}
```